

**ANOMALY DETECTION IN SURVEILLANCE
VIDEOS**

*Dissertation Submitted in partial fulfillment of the requirements for the
degree of*

**MASTER OF TECHNOLOGY
IN
ELECTRONICS & COMMUNICATION ENGINEERING
WITH SPECIALIZATION IN INTERNET OF THINGS**

By

AMIT ROY (225042002)

UNDER THE GUIDANCE OF

Dr. Nishant Jain & Dr. Vikas Baghel



**DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,
WAKNAGHAT, SOLAN-173234, HIMACHAL PRADESH**

May 2024

TABLE OF CONTENTS

CHAPTER	CAPTION	PAGE NO.
	DECLARATION.....	VI
	CERTIFICATE.....	VII
	ACKNOWLEDGEMENT.....	IX
	LIST OF ABBREVIATIONS.....	X
	LIST OF FIGURES.....	XI-XII
	LIST OF TABLES.....	XIII
	ABSTRACT.....	XIV
1	INTRODUCTION	1-11
1.1	Applications of Convolutional Neural Networks	2
1.1.1	Object Detection	2
1.1.2	Semantic Segmentation.....	5
1.1.3	Image Captioning	6
1.1.4	Audio Processing.....	7
1.1.5	Synthetic Data Generation.....	8
2	LITERATURE REVIEW	12-15
3	RESEARCH METHODOLOGY	16-33
3.1	CNN Model	17
3.1.1	Convolutional Layer.....	17
3.1.2	Pooling Layer.....	18
3.1.3	Fully Connected Layers.....	18
3.1.4	Working of CNN.....	19
3.1.5	ReLU.....	19

3.2	Different CNN Architectures	20
3.2.1	AlexNet	21
3.2.2	Inception V1 (GoogleNet)	21
3.2.3	Residual Neural Network (ResNet)	22
3.2.4	Visual Geometry group networks (VGGNet)	23
3.2.5	Inception-V3	24
3.2.6	InceptionResNetV2	25
3.2.7	NasNet.....	25
3.2.8	Dense Convolutional Network (DenseNet).....	26
3.2.9	Xception.....	27
3.2.10	MobileNets	28
3.2.11	EfficientNet	29
3.2.12	ConvNeXtXLarge.....	30
3.3	CNN-based Transfer Learning Techniques	31
4	EXPERIMENT AND RESULTS	34-39
4.1	Dataset for the Experiment.....	34
4.2	Methods Used for Performance Analysis.....	37
4.3	Analysis of Results.....	38
5	CONCLUSION	40
	REFERENCES	41-47
	PUBLICATIONS	48
	PLAGIARISM REPORT	49

DECLARATION

I hereby declare that the thesis work reported in this M.Tech Dissertation report entitled “**Anomaly Detection in Surveillance Videos**” submitted at **Jaypee University of Information Technology, Waknaghat, India** is an authentic record of my own work carried out under the guidance and supervision of Dr. Nishant Jain and Dr. Vikas Baghel. I have not submitted it elsewhere for the purpose of any other degree.

Signature of Student

Amit Roy
Enrollment No. 225042002



**JAYPEE UNIVERSITY OF INFORMATION
TECHNOLOGY**
WAKNAGHAT, P.O. – WAKNAGHAT,
TEHSIL – KANDAGHAT, DISTRICT – SOLAN (H.P.)
PIN – 173234 (INDIA) Phone Number- +91-1792-257999
(Established by H.P. State Legislature vide Act No. 14 of 2002)



CERTIFICATE

This is to certify that the work in the thesis entitled “**ANOMALY DETECTION IN SURVEILLANCE VIDEOS**” submitted by Amit Roy is a record of an original research work carried out by him under our supervision and guidance in partial fulfilment of the requirements for the award of the degree of Master of Technology in Electronics and Communication Engineering with specialization in Internet of Things in the department of Electronics and Communication Engineering in Jaypee University of Information Technology, Waknaghat, India. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Signature of Supervisor

Dr. Nishant Jain
Assistant Professor (Senior Grade)
Department of ECE, JUIT
Date:

Signature of Co-Supervisor

Dr. Vikas Baghel
Assistant Professor (Senior Grade)
Department of ECE, JUIT
Date:

Signature of Head of the Department

Prof. (Dr.) Rajiv Kumar
Head of the Department
Department of ECE, JUIT
Date:

ACKNOWLEDGEMENT

First, I would like to thank my parents for their continuous support and encouragement for my studies. It could never be possible for me to complete my second master's degree without their constant support. It is never possible for me to thank all the people who directly or indirectly helped me to reach wherever I am in with my career progress. I would like to express my gratitude to the head of my department Prof. Dr. Rajiv Kumar for his support during my degree program. Also, all the other faculties with whom I have engaged I give thanks to all of them for their suggestions, encouragement. I want to give a big thanks to both my supervisors Dr. Nishant Jain and Dr. Vikas Baghel for their encouraging comments and supports. My thanks to all the lab staffs for their silent support during my degree program in JUIT. I must have given thanks to all the library official and staffs as without their support I could never have such a great learning experiences where I spent most of my time. I also express my love for my only colleague Shubham for his long company to me for last two years. Before end, I again want to express my gratitude to my supervisor Dr. Nishant Jain without whom it could never been possible for me to explore the internal mechanism of Artificial Intelligence. He always used to have time for me considering his very busy schedules. His comments during the mentorship have helped me to shape my thought process. Each time I met with him one common comment was "you have to focus on basics, you have to know the fundamentals of the subject." When I am very close to complete my degree, I always carry this message with me. I express my gratitude to him from very deep inside of my heart.

I wish all of them a very happy and healthy life!

LIST OF ABBREVIATIONS

Serial No.	Abbreviation	Full Form
1.	CNN	Convolutional Neural Network
2.	LSTM	Long-Short term memory
3.	ReLU	Rectified Linear Unit
4.	MSE	Mean-square error
5.	GPU	Graphical Processing Unit
6.	BSVM	Bayesian Support Vector machine

LIST OF FIGURES

No. of Figures	Name of The Figures	Page No.
1.1	Example of object detection using SSD	3
1.2	Deep learning-based object detection detecting a person riding a horse and two plotted plants	4
1.3	Semantic Segmentation	5
1.4	Illustration of the CNN-RNN based image captioning framework	6
1.5	Audio classification application	7
1.6	Model takes a batch of pre-processed data and outputs class predictions	8
1.7	Synthetic data applications	9
1.8	Synthetic data applications	11
3.1	Flowchart of the Research Methodology	16
3.2	Architecture of the CNN model	17
3.3	ReLU Function	20
3.4	AlexNet Architecture	21
3.5	Architecture of Inception-V1	22
3.6	In Depth Architecture of ResNet50	23
3.7	Architecture of VGGNet	24
3.8	Inception V3 Architecture	24
3.9	Architecture of Inception-ResNet V2	25
3.10	Architecture of NasNet	26
3.11	Architecture of DenseNet	27
3.12	Architecture of Xception	28

3.13	MobileNet Architecture	29
3.14	Architecture of EfficientNet	30
3.15	Design of ConvNeXt	31
3.16	Working style of traditional machine learning and transfer learning	32
4.1	Sample images of training dataset	38

LIST OF TABLES

Serial No.	Table Name	Page No.
3.1	Comparison of different CNN models	33
4.1	Hyperparameters of CNN Models	34
4.2	Sample of the Dataset Labels	36
4.4	Predicted count with respect of actual count	38
4.4	Performance of CNN based Transfer Learning Techniques	39

ABSTRACT

Anomaly detection in surveillance videos is very important for ensuring the secure surveillance system. Different categories of anomaly events can occur in the surveillance videos, motion-based anomalies, appearance-based anomalies, audio-based anomalies are among them. Also, there are different types of anomaly detections like Intrusion detection, abandoned object detection, crowd behavior analysis and traffic incident detection. Accurately counting crowds is a big concern for the authorities when they organize major events in any part of the world. It is important to have a proper solution to this longtime problem with the techniques based on artificial intelligence. In this project work, for the proper estimation of crowd density different transfer learning techniques of the Convolutional neural networks have been used. A comparison among multiple transfer learning techniques like ResNet50, InceptionResNetV2, Xception, NasNet, VGG19, MobileNet, DenseNet, ConvNeXtXLarge, and EfficientNetV2L demonstrated using publicly available datasets for crowd counting. Parameters like mean square error (MSE) and Pearson's correlation coefficient (r) used for comparing among these CNN-based methods. In the project work, VGG19 has achieved the lowest value for MSE, i.e., 0.6 and highest value for Pearson r , i.e., 0.9.

CHAPTER - 1

INTRODUCTION

The rapid technological progress has appeared with a handful of opportunities for the common people. Public safety always gets priority for the government or the non-government places whenever there is any event organized. The use of advanced surveillance techniques are the only way to enhance safeguards with heightened security. Anomaly detection counts all the unusual events occurring in surveillance videos. An efficient anomaly detection technique plays a pivotal role in real-world people counting, efficiently identifying irregularities within large number of crowds. A large number of people lost their lives and got injury in events full of crowds due to managerial inefficiency. More than half of the world population will be urban by 2050 and for that societies need to develop smart crowd management systems [1]. Due to the increasing population, inefficient crowd management systems need to be upgraded using modern technological advantages.

On 3rd February 1954, over eight hundred people killed, and hundreds got injury in a stampede at the Maha Kumbhmela in Uttar Pradesh. On 23rd February and June 13, 1997, fire broke out and after that a forceful rash that followed left two hundred dead in Odisha's Baripada district and in a same kind of incident killed 59 people and injured over 100 people in Uphaar theatre in Mumbai respectively. In April, 2006, due to an electronics fault in a park Meerut killed over 100 people.

Another incident happened on August 2, 2008, nearly 162 devotees killed and approx. 50 people got injury in a stampede in the well-known pilgrimage Naina Devi temple in Bilaspur district of Himachal Pradesh. Another worst accident happened on April 10, 2016, where more than hundred lost lives and nearly 280 suffered injury in a powerful blast in a temple complex, in Kollam district. On January 1, 2022, also some people died and over a dozen suffered injury in a stampede at the renowned Mata Vaishno Devi shrine in Jammu and Kashmir [2].

The evolution of technology and the progress in deep neural networks enhances the chances of reducing the risks in crowded places. The usage of convolutional neural networks improves the people counting methods in recent techniques. In this project work, an attempt to improve the estimation of crowd counting with the trained models using multiple transfer learning techniques with a dataset of 2000 images have been taken. A comparison between the transfer learning

techniques help to identify the technique with minimum mean square error to design a better efficient model. Also, the findings of multiple literature related to people counting articulated the importance of the necessity of better people counting techniques. From enhancing crowd management strategies to preempting potential safety risks, this technique will improve accuracy and reduce the error percentage from all the other recently available techniques.

Image processing works with pictures which is computerized depiction of any image. Basically, pictures can be addressed by a cluster which is two layers organized in lines and sections or rows or columns. Basically, pictures are compromised with limited little components, called pixels. Viable data as indicated by the application is taken out by the system of image processing. Initial step of picture handling comprises of picture catching with some gadget. Image quality changing from color or gray scale additionally happens here and input images can be the result of any scanner or advanced camera. Processing comprises of image improvement and image rebuilding processes. Image upgrade or enhancement is indisputably the main part of image processing. Image segmentation dwells with the picture division and each part can be managed independently. If picture division is taken independently, it is convoluted. It is possible to analysis each part and can be performed freely. To portray or break down the image highlights can be separated in parts or from entire image. After processing images can be utilized in different applications. Some of the applications where it can be used are clinical imaging, interactive media and develop application based on computer vision.

Image processing has improved with the evolution of the neural networks more specifically Convolutional Neural Networks (CNN). CNN does the best performance for classifying an image. It consists of primarily three layers, convolutional, pooling and fully connected layers.

1.1 Applications of Convolutional Neural Networks

1.1.1 Object Detection

With CNN, we currently have modern models like R-CNN, Quick R-CNN, and Quicker R-CNN that are the transcendent pipeline for some item identification models conveyed in autonomous vehicles, facial detection, and that's very small part considering its giant capacity. Robotized driving

depends on CNNs precisely recognizing the presence of a sign or other item and simply decide. Deep learning-based object detection has different methods like Faster R-CNNs, You Only Look Once (YOLO), Single Shot Detectors (SSDs) object detection methods. Quicker R-CNNs are possibly the most "knew about" strategy for finding out object location utilizing deep learning; in any case, the procedure can be challenging to comprehend, difficult to execute, and testing to prepare. For unadulterated speed YOLO is the algorithm a lot quicker, equipped for handling 40-90 FPS on a Titan X GPU. Up to 155 frames per second is even possible with the super-fast YOLO variant. SSDs, initially created by Google, are a harmony between the two. The calculation is clearer than Quicker R-CNNs [3].

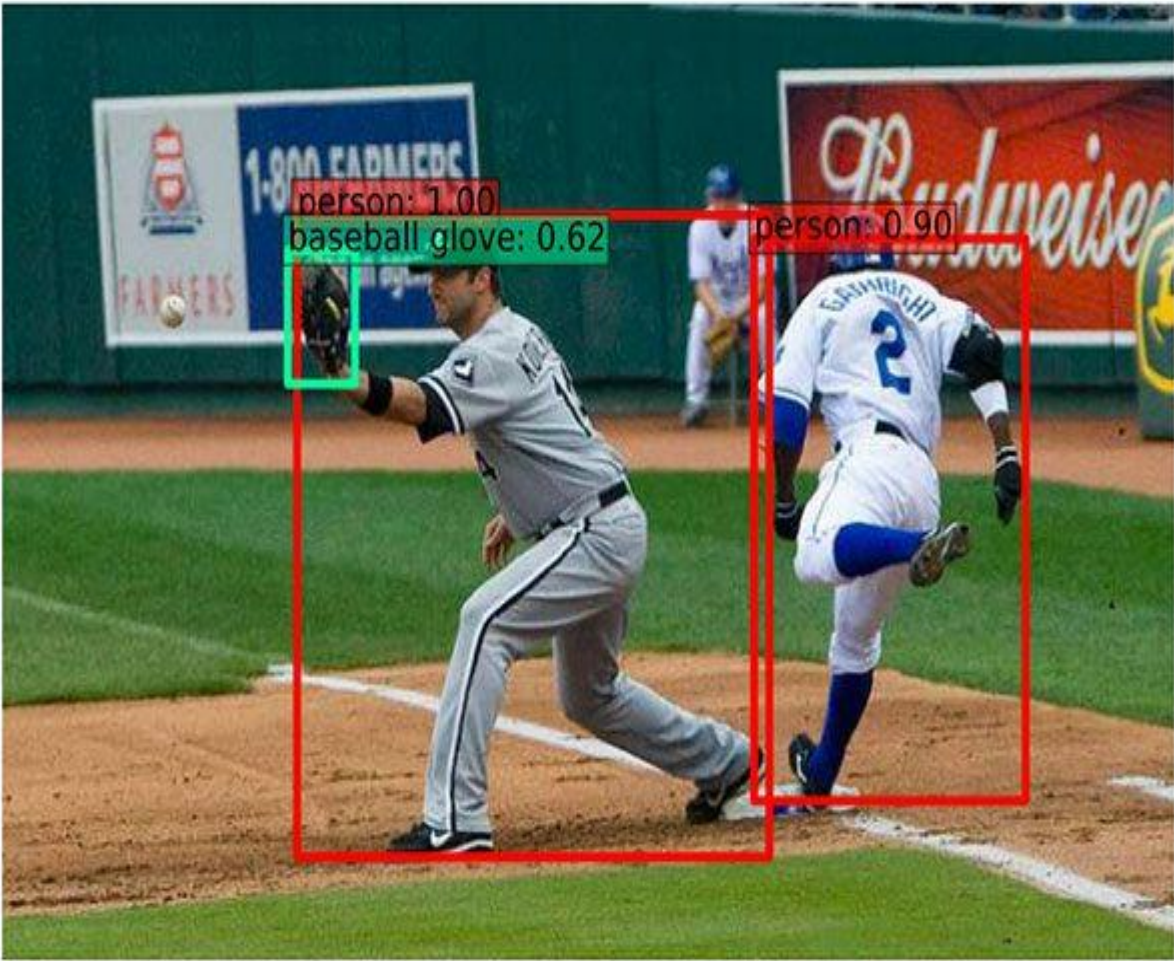


Figure 1.1: Example of object detection using SSD [3]



Figure1.2: Deep learning-based object detection detecting a person riding a horse and two potted plants. [3]

The capacity for deep learning to figure out how to distinguish and confine clouded objects is exhibited in the accompanying picture. Object discovery calculations need different and excellent information to ideally perform. A rich dataset library helps train more precise and versatile models, prepared for real world detection tasks.

1.1.2 Semantic Segmentation

In 2015, researchers from Hong Kong developed a CNN-based Deep Parsing Network to incorporate rich information into an image. Image segmentation methods range from straightforward, intuitive heuristic analysis to cutting-edge deep learning implementation. To identify object boundaries and background regions, conventional image segmentation algorithms process high-level visual features of each pixel, such as color or brightness. In AI, machine learning utilize clarified datasets, is used to prepare models to precisely order the kinds of items and areas of interest a picture contains. Image segmentation is a technique of computer vision that parcels a computerized picture into discrete gatherings of pixels — picture portions — to illuminate object discovery and related undertakings. Image segmentation makes it possible for faster and more advanced image processing by breaking up the intricate visual information in an image into precisely shaped segments. Semantic division is the most straightforward sort of picture division. Every pixel is given a semantic class by a semantic segmentation model, but no other context or information (such as objects) is produced. Semantic segmentation does not distinguish between things and stuff; it treats all pixels as stuff. A semantic segmentation model trained to identify specific classes on a city street, for instance, would produce segmentation masks indicating the boundaries and contours of each relevant class of thing (such as vehicles or light poles) and thing (such as roads and sidewalks), but it would not distinguish between multiple instances of the same class or count them. For instance, cars parked side by side could be considered one continuous "car" segment [4].

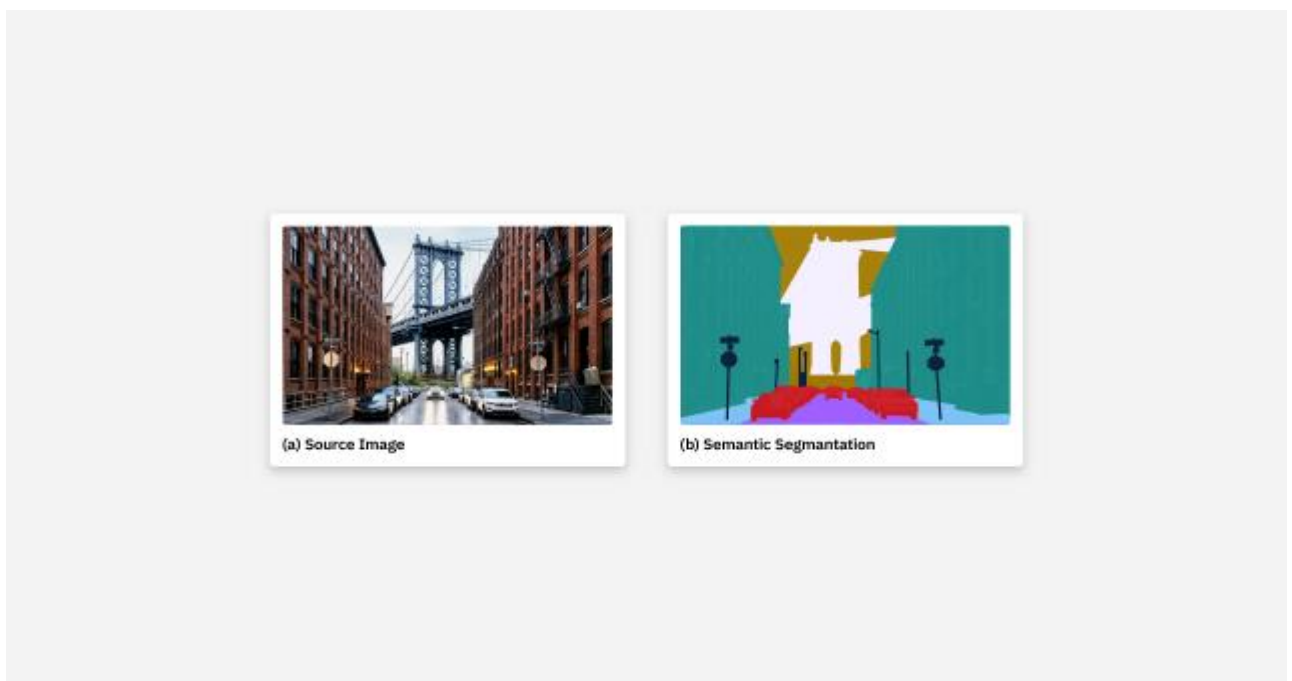


Figure 1.3: Semantic Segmentation [4]

1.1.3 Image Captioning

For writing the caption of the images and videos CNN used with recurrent neural network.

Medical Imaging: For detecting the presence or absence of cancer in the medical pathology reports CNN can be used.

Programmed picture subtitle age unites ongoing advances in natural language processing and computer vision. After advances in statistical language modeling and image recognition, image caption generation has emerged as a challenging and important research area. The option of inscriptions from pictures has different pragmatic advantages, going from helping the outwardly debilitated, to empowering the labeling of the millions of images that are uploaded to the Internet each day in an automated and cost-effective manner. The field additionally unites best in class models in Natural Language Processing and Computer Vision, two of the significant fields in Man-made reasoning. One of the principal challenges in the field of Picture Subtitling is overfitting the preparation information. This is on the grounds that the biggest datasets, like the Microsoft Common Objects in Context (MSCOCO) dataset, just have 160000 marked models, from which any hierarchical design should learn (a) a strong image representation, (b) a powerful covered up state LSTM representation to catch picture semantics and (c) language modeling for linguistically sound subtitle generation. The issue of overfitting shows itself in the retention of data sources and the utilization of comparative sounding subtitles for pictures which contrast in their subtitles. For instance, a picture of a man on a skateboard on a slope might get a similar subtitle has a picture of a man on a skateboard on a table [5].

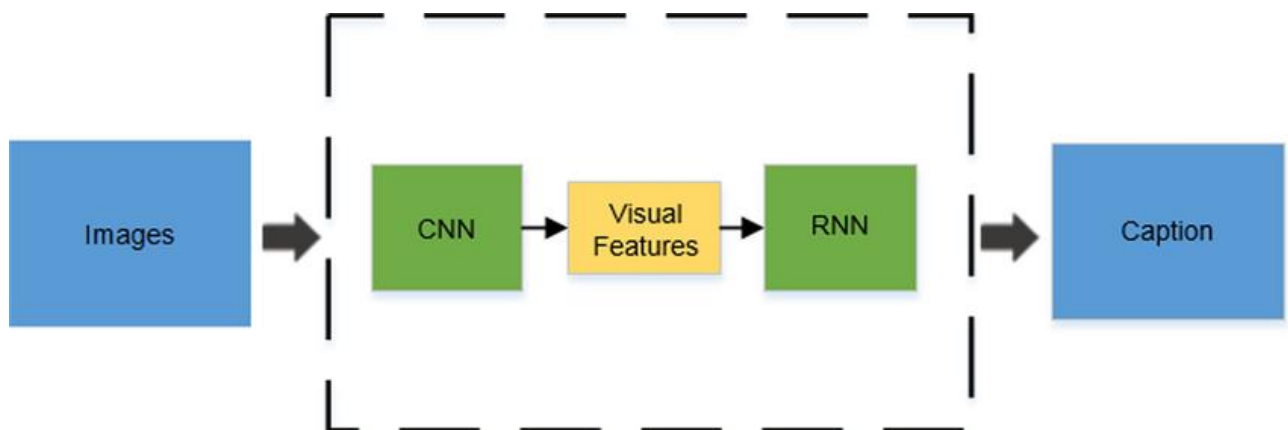


Figure 1.4: Illustration of the CNN-RNN based image captioning framework. [6]

1.1.4 Audio Processing

When someone speaks a word or used a phrase it is possible to use keyword detection in any device with a microphone to detect it. CNNs can be utilized to categories audio sounds into various classifications, for example, species identification considering bird or frog sounds. They can gain examples and elements from sound information, like how they process pictures, by changing over sound into visual representations like spectrograms. The use of CNNs for sound characterization, sound elements should be separated and changed into a format that looks like images. This should be possible by changing over sound signs into spectrograms, which are visual portrayals of recurrence spectra after some time. The extricated highlights are then taken care of into a CNN, which can figure out how to distinguish patterns and classify the sound signals accordingly.

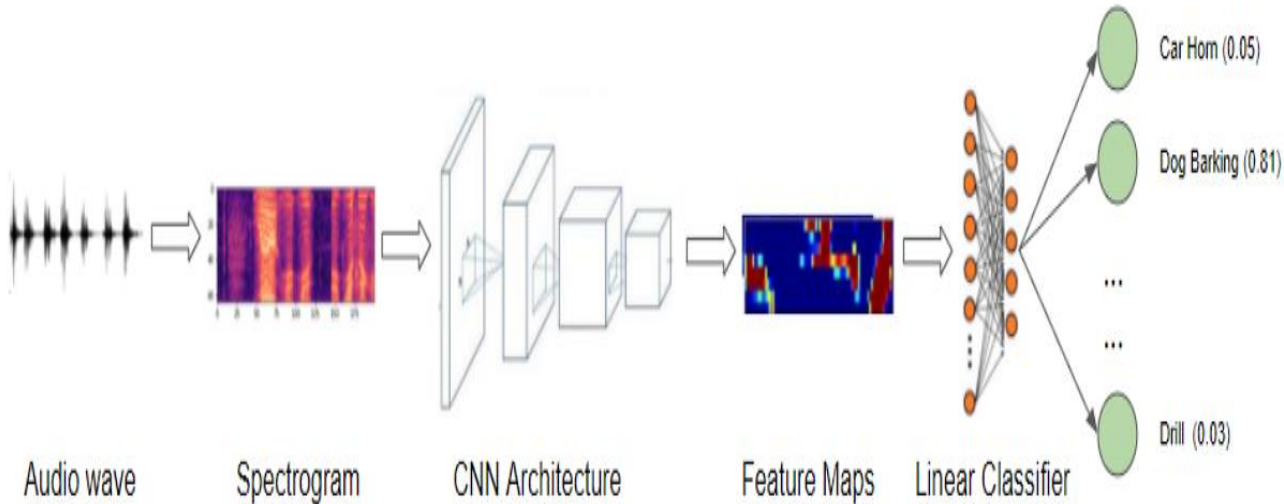


Figure 1.5: Audio classification application [7]

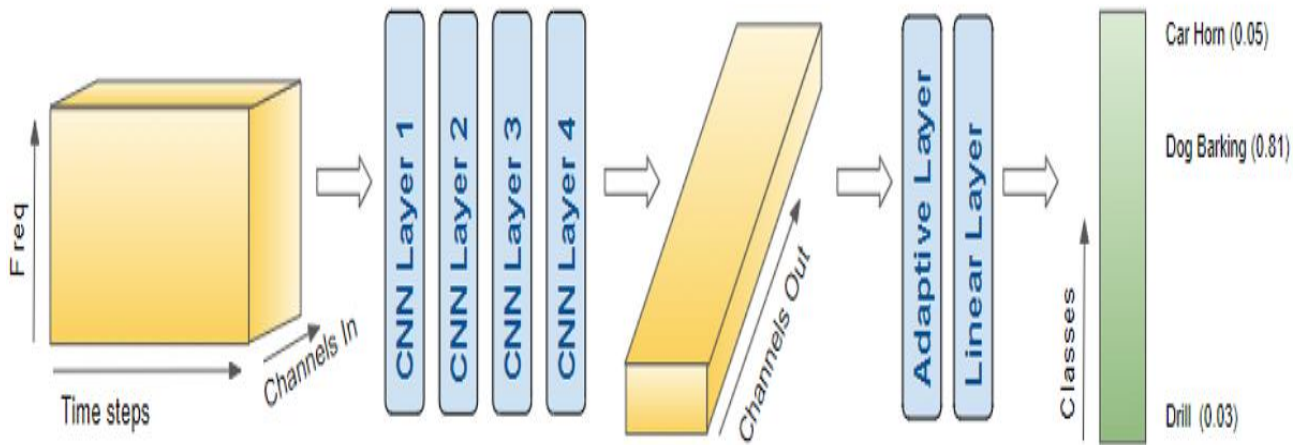


Figure 1.6: Model takes a batch of pre-processed data and outputs class predictions [7]

1.1.5 Synthetic Data Generation

Generative Adversarial Networks (GANs) can be used for producing new images using deep learning applications including face recognition and automated driving.

The idea of synthetic data generation arises as a promising elective that considers information sharing and use in manners that true information can't work with. Synthetic data are characterized as the falsely clarified data created by computer algorithms. Synthetic data is by and large characterized as falsely commented on data created by computer algorithms or simulations. Synthetic data offers many convincing benefits, making it an exceptionally engaging choice for many applications. This state-of-the-art innovation lessens the gamble of uncovering sensitive data, accordingly, guaranteeing client security and protection.

A generative adversarial network (GAN) that helps Visual Domain Adaptation by reducing the distance between embeddings in the learned feature space. This approach empowers semantic division across various spaces. The GAN utilizes a generator to extend highlights onto the image spaces, which the discriminator in this manner works on. Adversarial losses can be gotten from the discriminator's result. When compared to applying adversarial losses directly to the feature space, it has been demonstrated that applying them to the projected image space results in significantly superior performance. Synthetic data alone demonstrated adequate for recognizing faces in unconstrained settings. Trained machine learning system use tasks like landmark localization using syntheticdata.

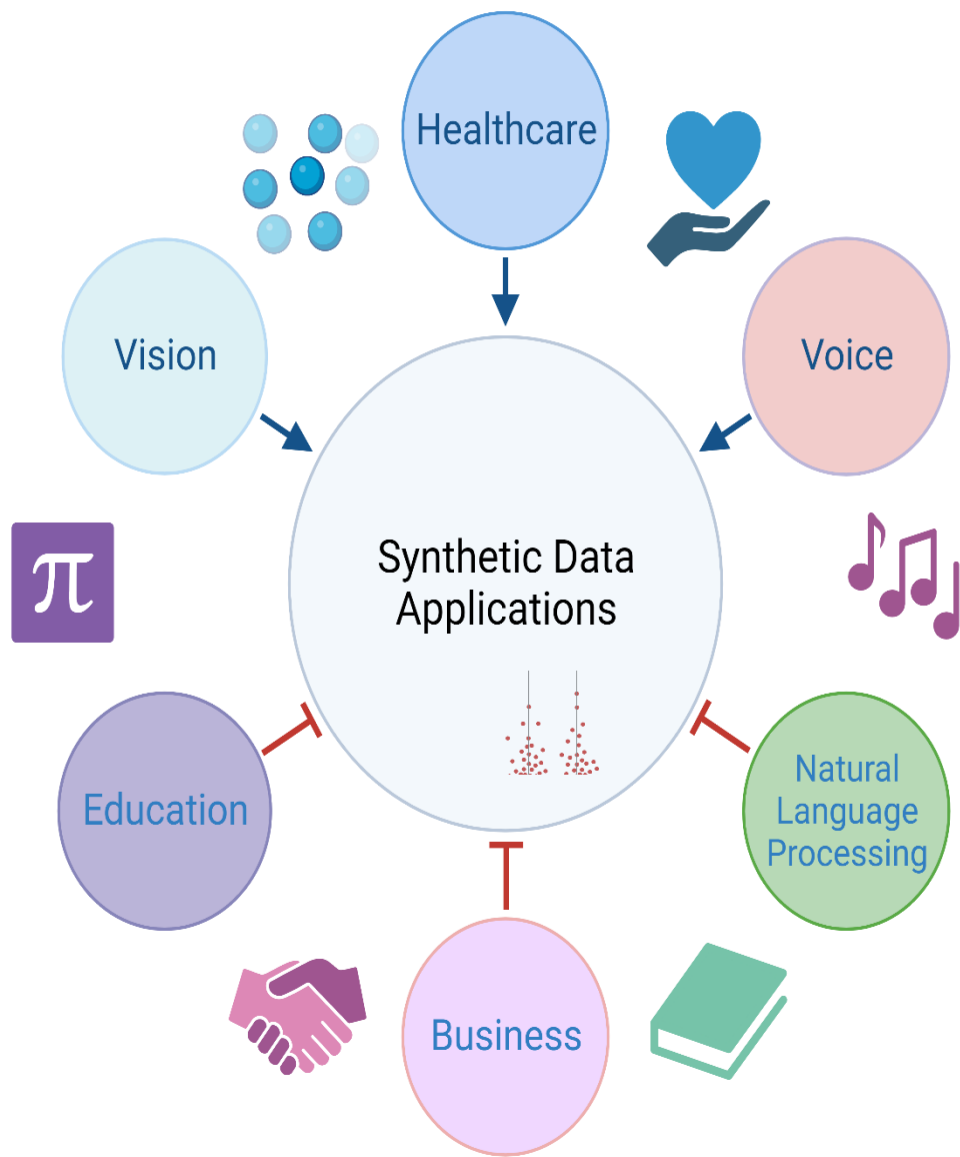


Figure 1.7: Synthetic data applications [8]

The field of synthetic voice is at the cutting edge of technology, and its development is accelerating at an alarming rate. With the approach of AI and profound picking up, making manufactured voices for different applications like video creation, computerized partners, and computer games has become simpler and more precise. This field is a convergence of assorted disciplines, including acoustics, phonetics, and signal processing. Analysts in this space consistently endeavor to work on synthetic voices' exactness and effortlessness. We can anticipate an increase in the use of synthetic voices in our day-to-day lives as technology develops, enhancing our experiences and assisting us in a variety of ways.

Synthetic information can likewise be applied to Text-to-Speech (TTS) to accomplish close human effortlessness. As an option in contrast to meager or restricted information, manufactured discourse (SynthASR) was produced for programmed discourse acknowledgment. The blend of weighted multi-style preprocessing, information enhancement, encoder freezing, and boundary regularization is additionally utilized to address catastrophic neglecting. Using this model, it is possible to train a wide variety of end-to-end (E2E) automatic speech recognition (ASR) models by employing this novel model, thereby reducing the requirement for production data and the associated costs.

In the field of natural language processing (NLP), a wide range of deep generative models have been developed with synthetic data. A large number of techniques and models have delineated the capacities of AI in sorting, directing, sifting, and looking for pertinent data across different spaces. The BLEURT model was proposed, which models human decisions utilizing a set number of possibly one-sided preparing models in view of BERT. The specialists utilized large number of synthetic examples to foster an inventive pre-preparing plan, supporting the model's capacity to sum up. Trial results show that BLEURT outperforms its partners on both the WebNLG adverse dataset and the WMT Measurements, featuring its adequacy in NLP undertakings.

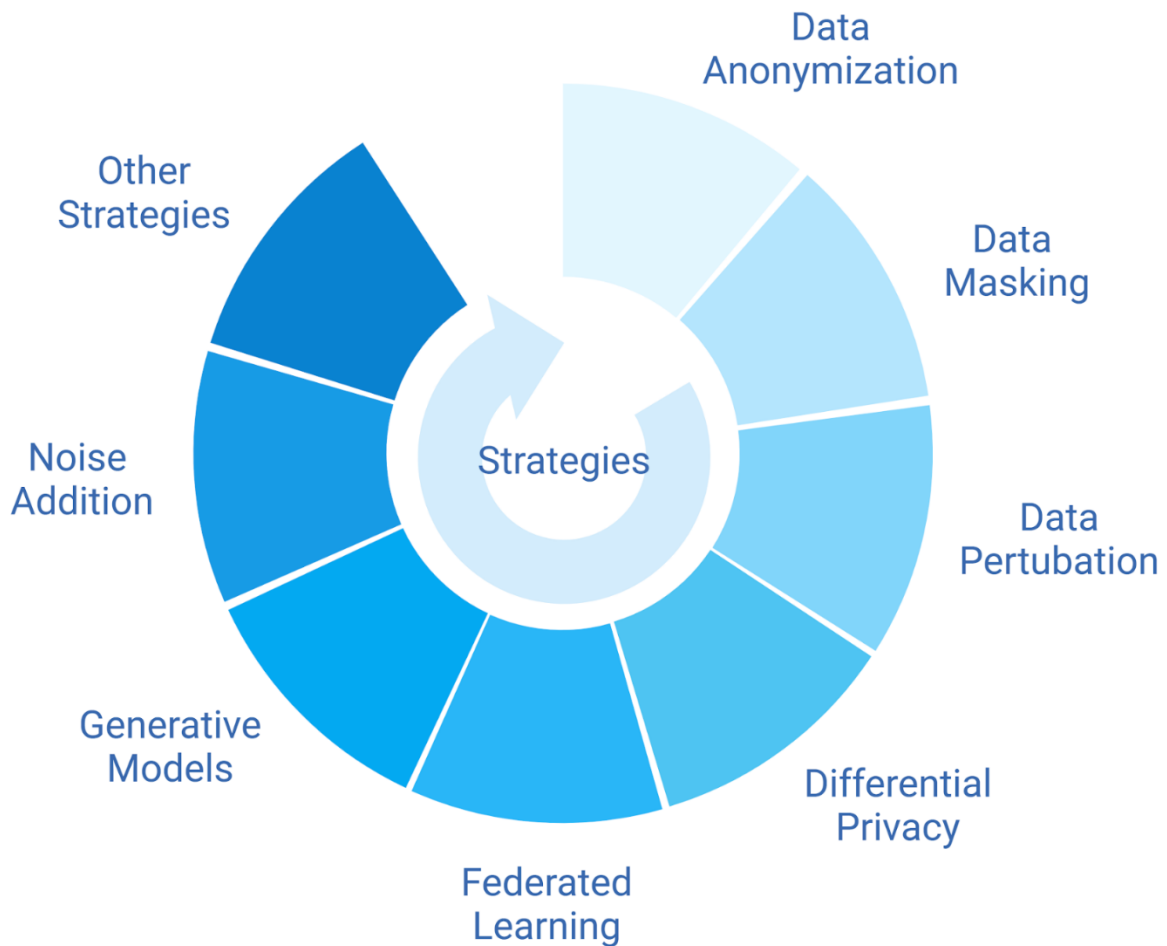


Figure 1.8: Synthetic data applications [8]

The utilization of synthetic information is turning into a suitable option in contrast to preparing models with genuine information because of advances in reproductions and generative models. Various challenges should be defeated to accomplish elite execution. These incorporate the absence of standard devices, the contrast between manufactured and genuine information, and how much AI calculations can do to really take advantage of blemished synthetic information. However, this arising approach is flawed now, with models, measurements, and advances developing, synthetic information will have a greater effect from here on out [8].

CHAPTER - 2

LITERATURE REVIEW

Researchers are working for a long time to improve the crowd counting methodology to enhance the accuracy in the realm of anomaly detection. The evolution of the neural networks helps to implement the deep learning techniques for accurately tracking the movement of individuals within largely crowded events. The World Health Organization (WHO) draws the definition of crowded events where a significant number of participants gathered planned or unplanned for any occurrence which strain the city or nation hosting events planning [9]. In this increasingly crowded world, specifically in third world countries it is important to have a smart system for counting crowds from moving image frames. Crowd management monitoring systems will prevent the risks, improve public safety conditions, and will even prevent possible disasters which may take human lives.

Anomaly detection in surveillance systems is critical for the prevention of possible disasters and in some cases for security purposes [10]. In 2009, authors in this paper [11] used the texture and foreground features for generating the low-level information from the images and find out the number of people from the relationship of the crowd and the extracted features from the input images.

Crowd management systems use different methods for planning and managing mass gathering events. Measuring the public safety, event managing, planning, predicts and prevents unwanted incidents and helps to prepare plans for emergencies are the focus of the crowd management system. Crowd monitoring helps to estimate crowd dynamics, detection and prediction of the amount of risks associated with and tracking and supporting virtual simulation of crowd behavior for the development of automated systems [12]. In 2013, a model proposed by Idrees et al., which introduced the use of Fourier analysis and SIFT (Scale invariant feature transform) to collect the low-level information of important points to get the idea of crowd density. According to this paper [13], anomaly detection has many different categories among them accurately identifying crowd movements is one where behavioral abnormality in crowded locations emerge as crowd commotion. Detecting anomalies in the images targets to identify and categorize abnormalities in datasets. For analyzing the textual data, anomalies can be captured by data plotting and each data point greater or less than other data in the datasets are known as anomalies or outliers. In the video or image data also, it is possible to identify the anomaly critically by analyzing and observing the behavior or

object patterns in the concerned area. The abnormal behavior of the object helps to identify anomaly in object. Anomaly detection detects the patterns that do not correspond to expectations. Different abnormal activities like congestion, avoiding the usage of the pedestrian path, avoiding areas not specified for standing and hampering the forward motion of other pilgrims, refrain from creating chaos by running and causing commotion at the entry and exit points and transit stations. As violent behavior also abnormal behavior, which affects the people, it can also be possible to detect the abnormalities through a smart system that can control the safety of the environment and also limits the possibility of violations that happen in different incidents [14].

This paper [15] used 3D CNN-based models to detect accidents to address different weather and lighting conditions in different situations. In the testing phase two functions known as loss functions are used with and without optical flow patterns to extract the feature of conditions with illumination from the background scenes. This model was tested with real-world traffic videos.

Before that author in this paper [16] used Gaussian Mixture Model to detect vehicles and mean of shift algorithms to track vehicles. This approach follows humanistic ideas and model collaboration between the vehicles for discovery of mishaps on road. The different techniques of deep learning addressed visual monitoring for road traffic and road surface monitoring systems.

Authors [17] used a multiple scaling feature of a fusion network with single column kernel convolution to extract the features and solve the distortion problem of large number of dense crowds. In this paper [18] they introduced the inversely attentional module to distinguish the presence of the crowds from proposed method divided the surveillance videos into segments during training and formulated anomaly detection as a problem of regression. They used deep multiple instance learning techniques to have higher anomaly scores comparing between anomalous video segments and normal segments.

The authors of this paper [19] articulated the progress of obstacles detecting on the road by applying an autoencoder with semantic segmentation. In the encoding part autoencoder contains image generator which is semantic, and another one is photographic as decoder. This is an unsupervised application where the model is trained on common road scenes only. This paper [20] applies deep learning models in the view of CNN, long short-term memory networks, and supply processing models to distinguish potholes and street surface characterization.

According to the paper [21] the 3D-CNN and LSTM models are utilized to keep up with correlation between consecutive images utilizing 3D-resNets design. The examinations proposed the technique has a phenomenal exhibition in unusual conduct acknowledgement on some difficult datasets. They had fostered a programmed strange conduct location arrangement of recordings in view of VGGNet, and BSVM, utilizing move learning strategies to distinguish unusual occasions. The outcomes showed that the VGGNet-19 acquired preferable exactness over different procedures, with a normal of 97.44 percent. A new fully connected convolutional neural networks (FCNs) engineering framework for worldwide strange conduct discovered global behavior detection [22].

In this project work the problem of crowd counting in the surveillance videos articulated as an anomaly detection problem. The problem of crowd counting occurs for various reasons like occlusion, perspective distortions which impacts the accuracy of the counting people. In this paper [23] authors proposed a self-attention mechanism-based counting method which solved the problem of crowd occlusion but a problem of perspective distortion that impacts the accuracy of the crowd counting in a negative way still exist. For getting the proper accuracy from the surveillance videos, processing the images is the most important part of the research. In [24] proposed a mechanism based on attention which is generating image analysis sentences for images by using a mathematical model. After developing the model further, the calculation process of this model has been enhanced with the attention convolution module. But with the increasing complexity the problem of accurate crowd counting is still unsolved.

Here authors [25] presented new ways to count crowd which is contextually aware pyramid attention network which extract high quality features based on context and dealt dependent on space and channels, but the extracted interference factors, which will impact counting accuracy, was not considered that the extracted features are excessively rich. This paper used image data as an input in a CNN model. The model will count the number of the people present in a particular image. Once fine-tuned the model with the multiple transfer learning techniques, it gets better results with the mean square error and pearson r to increase the people counting in the images. The transfer learning models like DenseNet, ConvNeXtXLarge, EfficientNetV2L, Inception, MobileNet, NesNet, VGG19, Xception used for the experiments to get the mean square error and pearson r on crowd counting.

The datasets used for training the model were taken in a mall and it is composed of RGB image frames in a video. There is 2000 RGB images of 480*640 pixels at three channels of the same places recorded in a webcam and this dataset doesn't have a similar number of people on every frame, which is an issue of people counting [26]. The model used CNN based classification approach to extract different features from the images, segmentation, background eliminations.

CHAPTER - 3

RESEARCH METHODOLOGY

In this project work image will be used as input data where the model will be used for counting the people present in that image. If the counting in the image don't cross the threshold or remain smaller or equal it will get back for the processing of the next image but if it crosses the threshold, if the count in the image greater than the threshold it will detect the presence of the anomaly in the system and indicate it by displaying. After displaying the presence of the anomaly in the image it will again go back to process for the next image. The whole process of the counting of the people will be done by the CNN model using the transfer learning techniques. An idea of the working procedure of the model is give below with the flowchart.

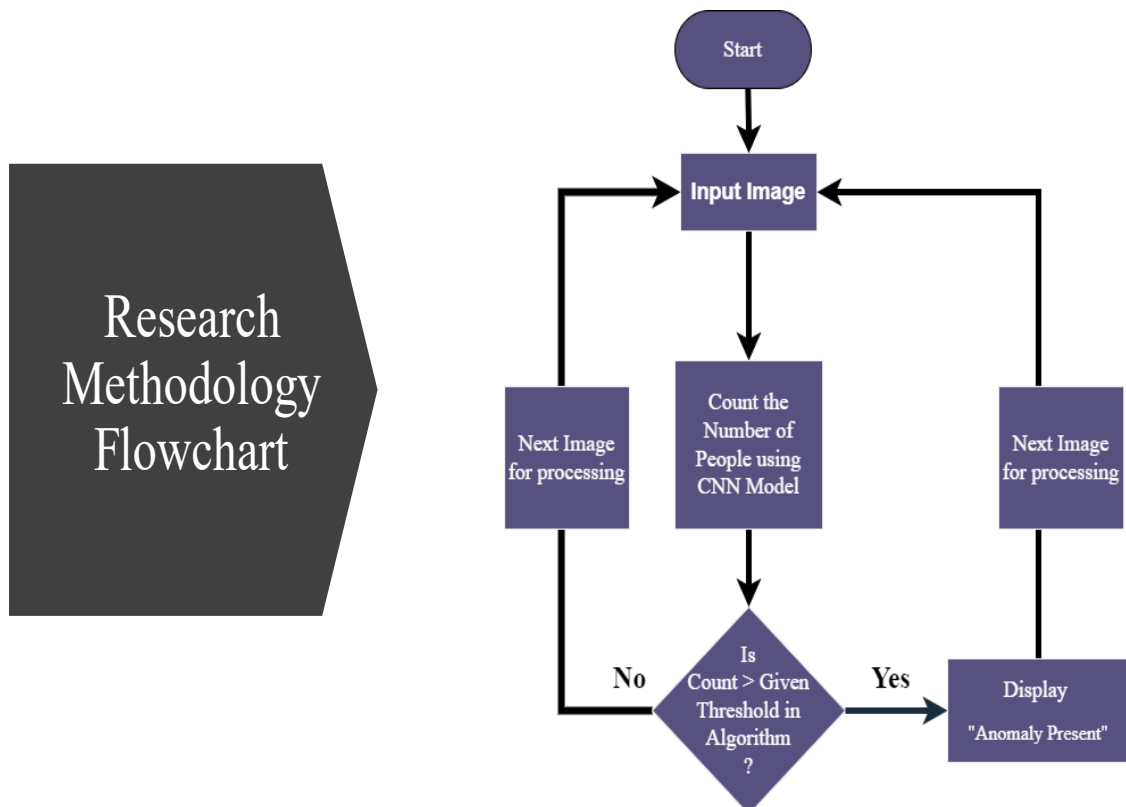


Figure 3.1 - Flowchart of the Research Methodology

3.1 CNN Model

Convolutional neural network (CNN) is a kind of deep learning model that is especially appropriate for dissecting visual information. It is roused of the creature visual cortex association and is intended in learning spatial ordered progressions of elements naturally and adaptively, from significantly low-level examples.

A CNN model basically consists of several layers, which can be broadly categorized in the three main groups known as convolutional layers, pooling layers and fully connected layers. As the input given in this layer, the complexity of the CNN increases with that and allow it in a supersessive way identify larger portions of an image and more robust features.

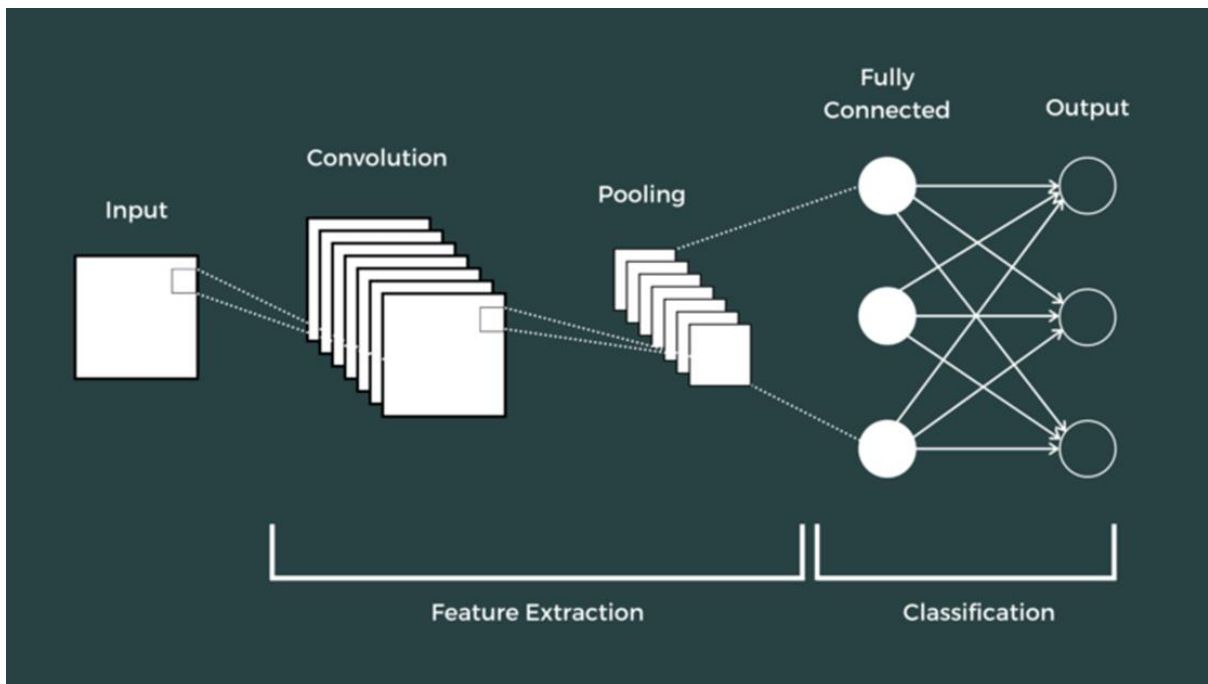


Figure 3.2 - Architecture of the CNN model

3.1.1 Convolutional Layer

Convolutional layer is the base of the CNN and it consists of a large number of filters which impact the depth of the output. This is the layer where most the computations occur. These layers are consisting of filters or kernels and weights to adjust the biases. The process starts by sliding the kernel or filters over the image's width and height, and over iterations multiple times it sweeps

across the entire images. All the positions it moves, it calculates image pixel values with a dot product between kernel's weights with respect of each kernel. This convolved the input image features and transforms the input image into a set of feature maps.

The equation use for the expression of convolutional layer function is:

$$f(X) = \sum_{k=0}^c W_k * X_k + b$$

W is the weight vector; k is total number of nodes and * denotes the operator of convolution networks [27].

3.1.2 Pooling Layer

Pooling layers of CNN model is the critical component that connected with the convolutional layer. Following the convolutional layer, it reduces spatial dimensions of the feature maps to improve efficiency and accuracy and used for dimensionality reduction. This layer is also known as dimensionality reduction layer. Pooling layers help to enhance complexity and limit inaccurate predictions and poor performance. There are two main pooling layers we use for; one is Max pooling, and another is Avg. pooling. This is done through down sampling, which captures an increasing larger field of view and increases model efficiency through reducing the number of learning parameters.

3.1.3 Fully Connected Layers

The following comes the fully connected layers where every hub in the result layer interfaces straightforwardly to the hub in the successive layer. These fully connected layer also known as the output layer. Likewise, classification, it maps the extracted features into final output. These layer takes output from the previous layers, convolutional layers and pooling layers and uses it for the final decision for classifying and object in an image. FC layers are applied after the series of resulting convolution and pooling layers to carry out the result leveling into a single vector.

3.1.4 Working of CNN

CNN has layer series where each of these layers finds out different features of an input image. Based on the complexity of the purpose of the usage, CNN models can contain dozens, hundreds or more than thousands of layers. It detects patterns depending on the building of the outputs of the previous layers. The process starts with sliding a designed filter to detect some known features in the image. For enabling the CNN model to gradually build a following representation of the images feature map serves as the next layer input.

3.1.5 ReLU

Rectified Linear Units save data about relative forces as data goes through numerous layers of element indicators. A downside of giving each duplicate a predisposition that contrasts by a proper offset is that the calculated capability should be utilized ordinarily to get the probabilities expected for inspecting a number worth accurately. The quick estimation of the sampled value of the rectified linear value is not constrained to be integer [28]. The activation function ReLU is used in the neural networks to get the output of the input directly if it is positive, otherwise zero if it is negative. This function is simple to compute, and it accelerates the training speed by avoiding the vanishing gradient problem.

The weight vectors balanced during the training period through the algorithmic backpropagation to acquire the feature maps convolving the input images and weights. Each convolutional operation in a layer applies a Rectified Linear Unit (ReLU) function to introduce non-linearity to the mode in the feature map.

The ReLU function is denoted by:

$$ReLU(X) = \left\{ \begin{array}{ll} 0, & \text{if } X < 0 \\ X, & \text{otherwise} \end{array} \right\}$$

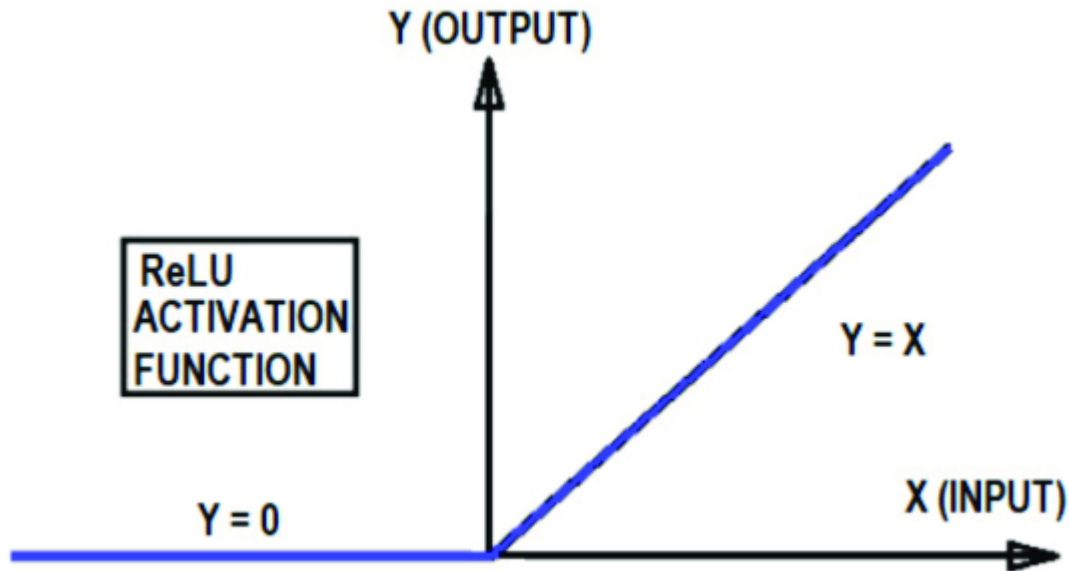


Figure 3.3 - ReLU Function [29]

Convolutional neural networks are strong sorts of artificial neural network that are especially appropriate for recognizing images and processing undertakings. They are inspired by the construction of the human visual cortex and have a various leveled design that permits them to gain and concentrate highlights from images at various scales. CNNs have been demonstrated to be exceptionally powerful in many applications, including classifying images, object detecting, image segmenting, and image generating.

3.2 Different CNN Architectures

Different CNN models trained on dataset consisting of very large number of images are LeNet [30], AlexNet [31], GoogleNet [32], VGGNet [33], ResNet [34], Xception [35], NasNet [36], MobileNet [37], DenseNet [38], EfficientNet [39], ConvNeXtXLarge [40]. Details of some of the above CNN models are discussed below:

3.2.1 AlexNet

LeCun et. al, proposed primary LeNet architecture was very similar to AlexNet, which was introduced in 2012 [30]. In correlation with the underlying one, the AlexNet design is not a lot further remembering beneficial channels for each of the layer. It contains variable convolutional channels of multiple sizes 11×11 , 5×5 , 3×3 , and then max pooling and ReLU activation are appended after each convolutional layer [31].

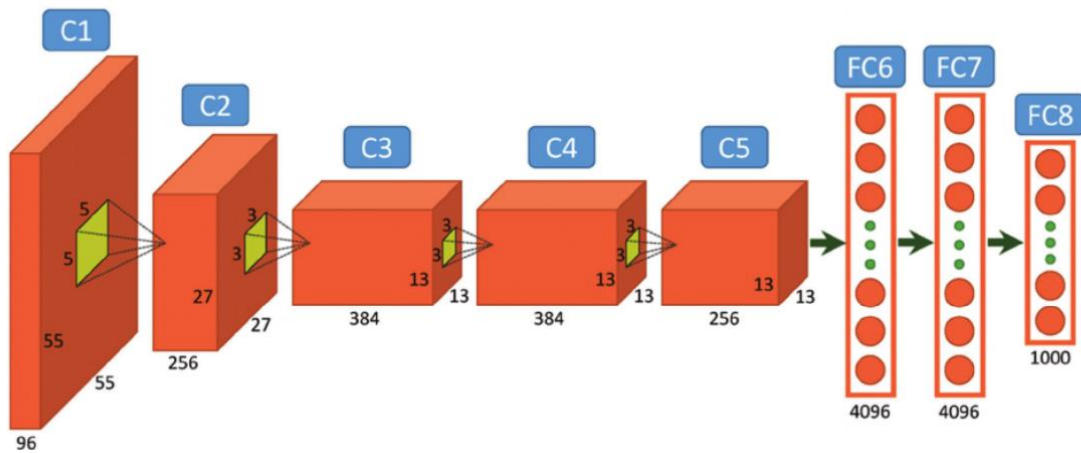


Figure 3.4 - AlexNet Architecture [42]

AlexNet was quick to execute the ReLU activation capability to propel the preparation of speed when expanding the network execution. The size of network decreased using pooling layer and the dropout layer to diminish overfitting. AlexNet lessens the learning rate during the preparation interaction when the precision esteem levels. To beat this issue, GoogleNet appeared in 2014.

3.2.2 Inception V1 (GoogleNet)

GoogleNet [32] was introduced in 2014 with an advantage of giving a limited blunder rate comparative with other open CNN networks. The detriment of this design is that it immerses the network precision while growing its profundity, consequently, ResNet engineering showed up in 2015 that utilizes the idea of skip associations while saving the profundity of the model.

Inception-V1 (GoogLeNet)

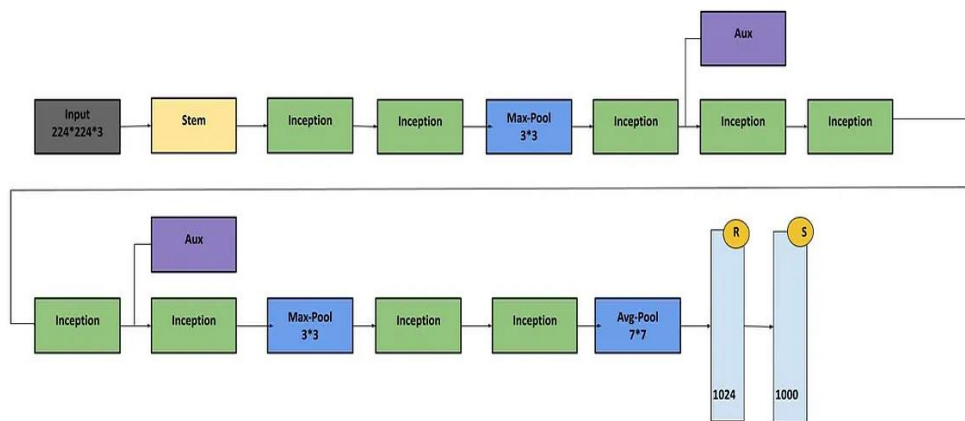


Figure 3.5 - Architecture of Inception-V1[43]

3.2.3 Residual Neural Network (ResNet)

Since its presentation in 2015, ResNet [34] has become increasingly popular for applications in the fields of facial recognition, object recognition, and image recognition due to its potent representation capability. ResNet gives great network execution, yet its design is to some degree perplexing and complex. Yet, as various CNN varieties showed up in 2015, VggNet is one of them which prominently has less complex execution as well as extends the networks profundity.

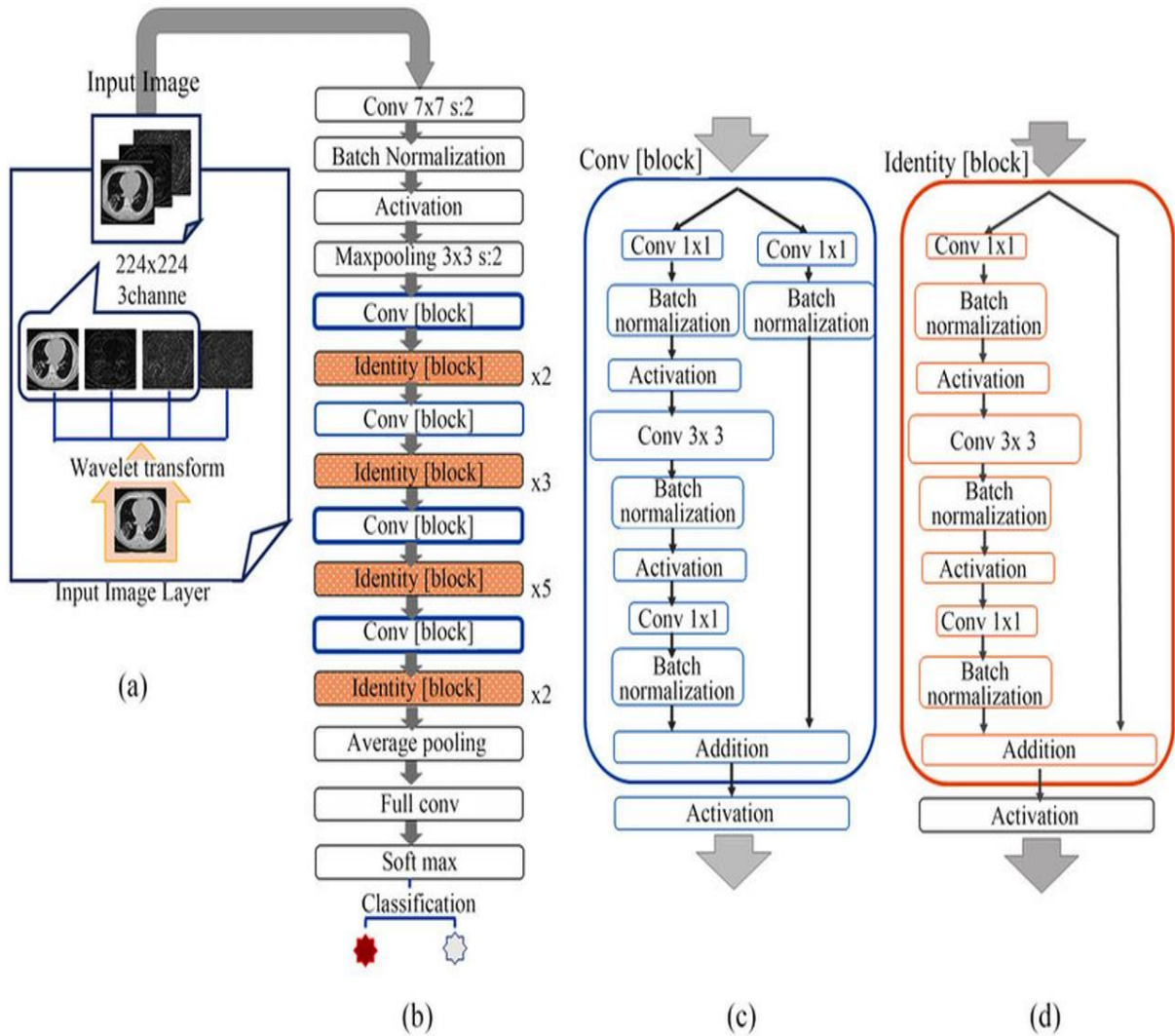


Figure 3.6 – In Depth Architecture of ResNet50 [44]

3.2.4 Visual Geometry group networks (VGGNet)

VGGNet [33] is a diverse deep neural network design that was introduced in 2015 because of its basic execution and further developed network profundity. VggNet contains an enormous number of network boundaries using a higher extra room of 500 MB. VGGNet has a limitation of slow preparation alongside higher extra room necessities making its network dreary. This obstruction is tended to by one more subsidiary of Inception, the Inception V3 model.

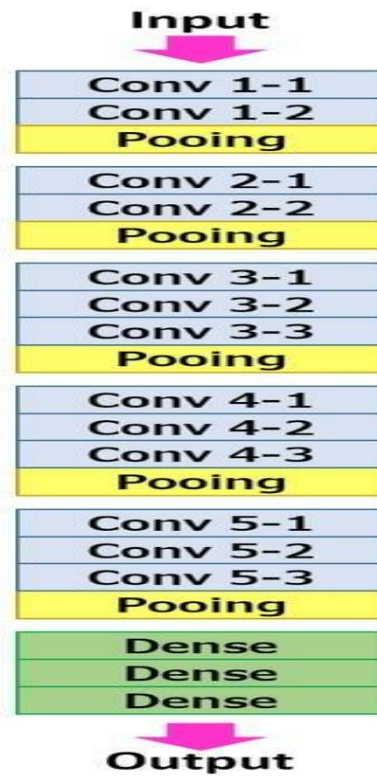


Figure 3.7 - Architecture of VGGNet [45]

3.2.5 Inception-V3

V3 is a profound deep neural network planned to sort 1000 item classifications [41]. A wide assortment of pictures is utilized to prepare the model and keeping up with that preparing information, the model can be retrained for a more modest dataset. This advantage of Inception-V3 CNN model diminishes the requirement for broad preparation bringing about higher characterization exactness alongside at least computational time.

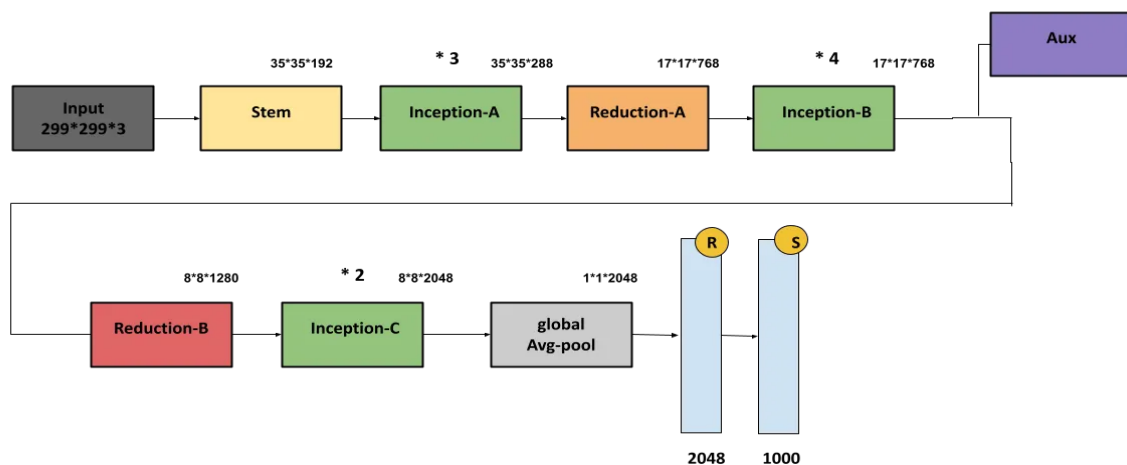


Figure 3.8 - Inception V3 Architecture [46]

3.2.6 InceptionResNetV2

InceptionResNetV2 is the remaining form of the Inception Network. This paper [41] utilized less expensive Initiation blocks than the first Inception. Each of the blocks are trailed by the filter of expansion layer (1×1 convolution without activation) which is utilized for increasing the dimension of the channel bank before the expansion to match the profundity of the information which used to make up for the dimensionality decrease instigated by the Inception block.

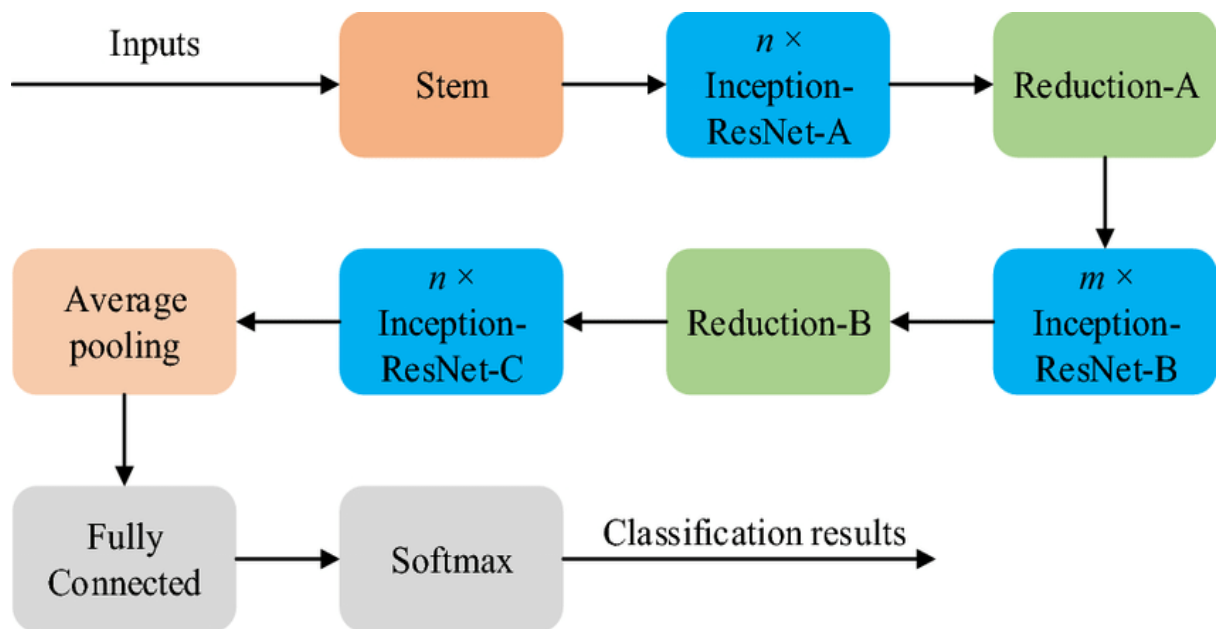


Figure 3.9 - Architecture of Inception-ResNet V2 [47]

3.2.7 NasNet

Neural Architecture Search network prepared on the more modest CIFAR-10 dataset, and afterward moved the learned architecture to ImageNet. NasNet architecture created the best convolutional layer on the CIFAR-10 dataset and afterward applied those layers on the ImageNet dataset by stacking together more layers each with their own boundaries to plan a convolutional design. By avoiding the depth of the network and the size of the input images, a search space designed to maintain the architecture's independence allowed for the transfer of learned architecture [36].

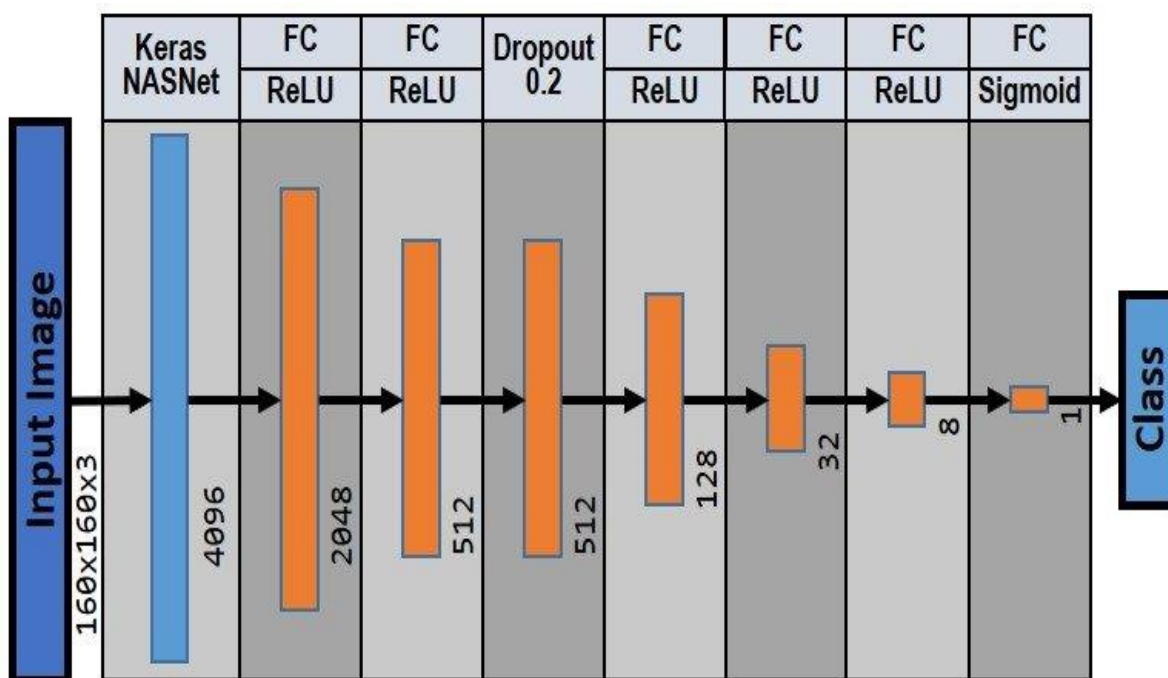


Figure 3.10 - Architecture of NasNet [48]

3.2.8 Dense Convolutional Network (DenseNet)

DenseNet associates with layer to layer in a feed-forward network. The feature maps of the layers that came before it is used as inputs for the layers, and the feature maps from it are used as inputs for all the layers that come after it. DenseNets enjoy a few convincing upper hands over the other deep neural networks. DenseNet eases the evaporating angle issue, reinforces include engendering, empowers highlight reuse and significantly diminishes the quantity of boundaries. The architecture of the DenseNet layer is extremely limited, they just have 1 filter for each layer while adding just a little arrangement of element guides to the aggregate information on the network and keeping the component maps unaltered where the last classifier pursues a choice in light of all component maps in the network [38].

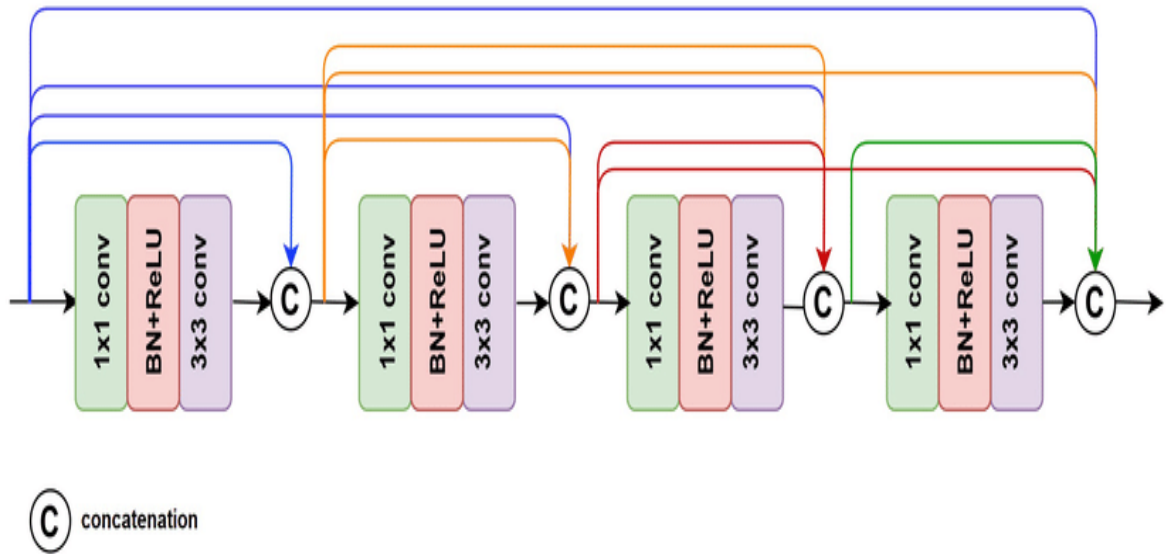


Figure 3.11 - Architecture of DenseNet [49]

3.2.9 Xception

The architecture of Xception has 36 convolutional layers which shape the network's component extraction base. These layers of 36 convolution are organized into 14 modules and has direct residual associations around them, aside from the first and last modules. The Xception architecture has the comparative boundary considering Inception V3 while contrasting both Xception shows little acquires in order execution on the dataset of ImageNet and huge additions on the JFT dataset [35].

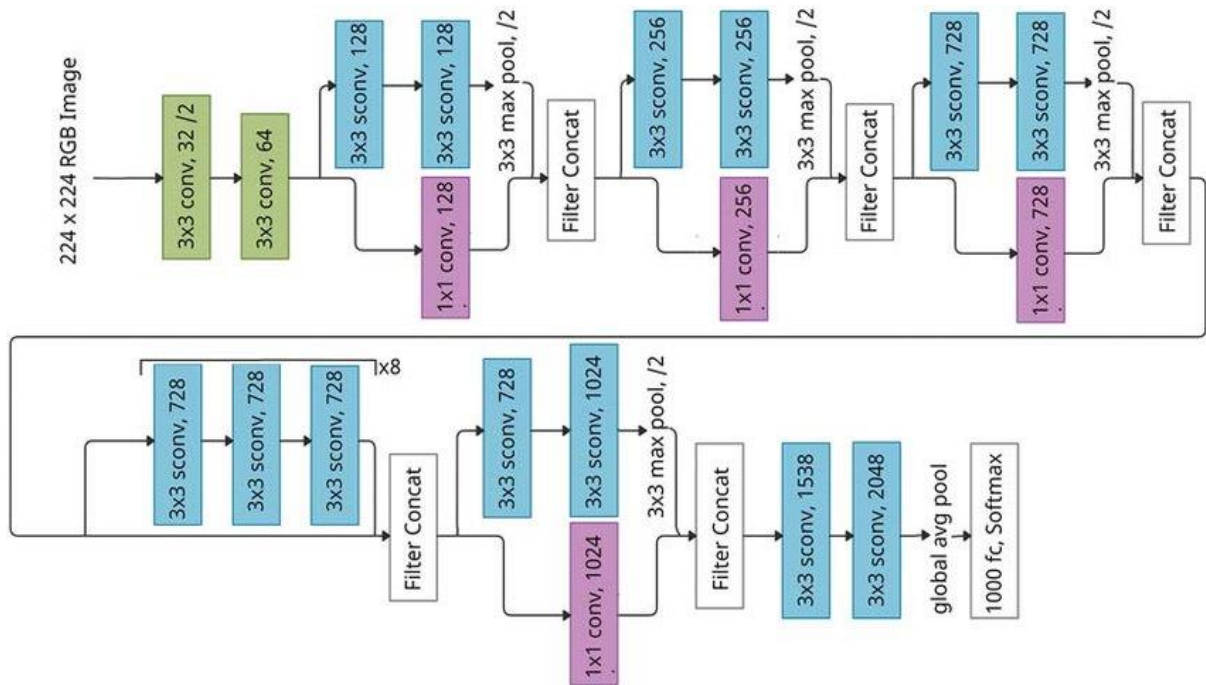


Figure 3.12 - Architecture of Xception [50]

3.2.10 MobileNets

MobileNets depend on a smoothed-out architecture that utilizes profundity wise distinct convolutions to construct light weight deep neural networks. MobileNets hence involved the Inception models to decrease the calculation in the initial not many layers. This architecture is a type of factorized convolutional which factorized a standard convolution into a 1 * 1 depth wise convolution called a pointwise convolution. For the MobileNets a single channel depth wise convolution applies to each information channel. The pointwise convolution then applies a 1 * 1 convolution to consolidate the results of the depth wise convolution. This standard convolution channels and consolidates the contribution to another newly arrangement of results in a single step. The depth wise separable convolution then divides this into two layers, one for combining and one for filtering. This has the impact of radically decreasing the computational power and the model size [37].

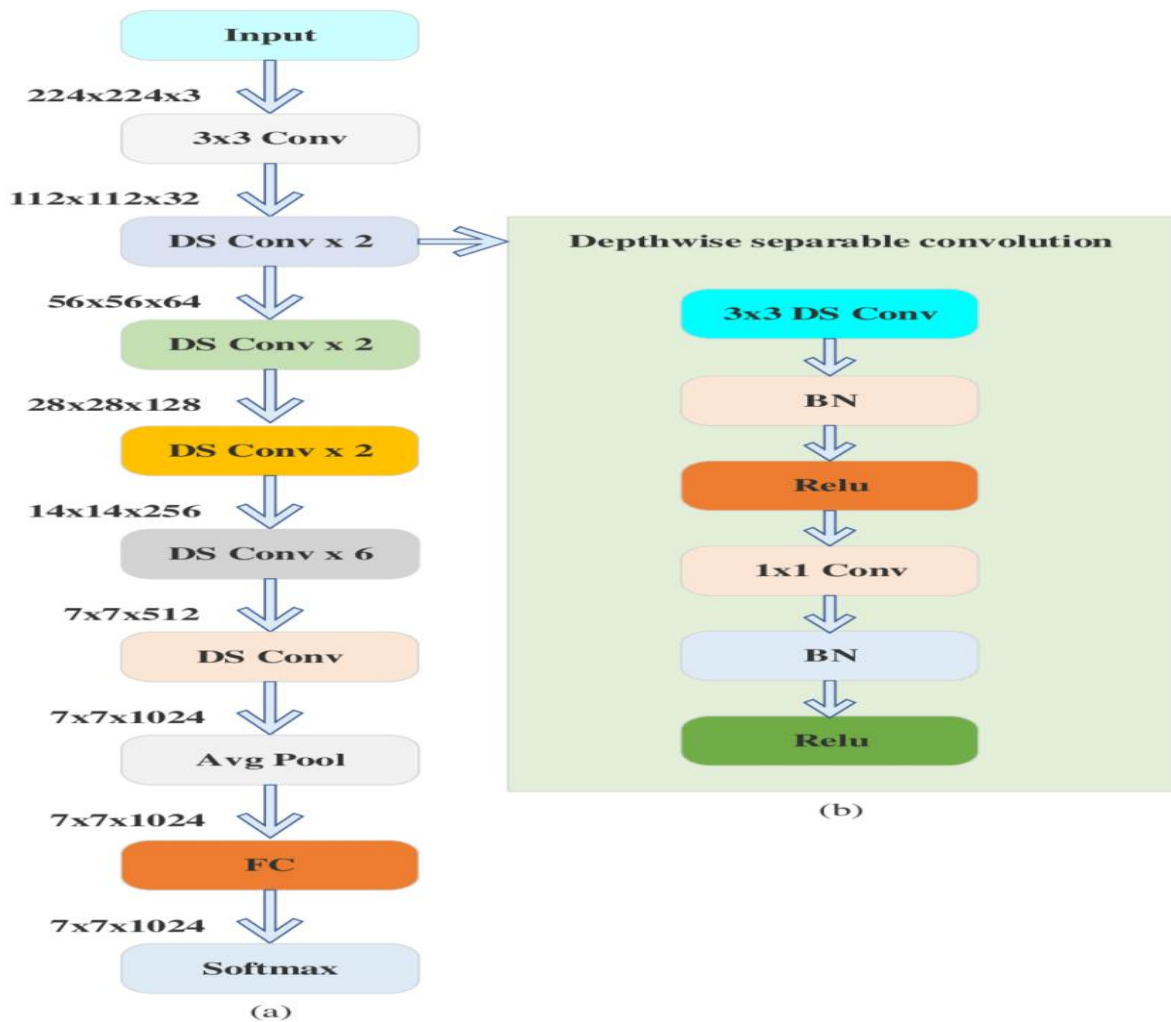


Figure 3.13 - MobileNet Architecture [51]

3.2.11 EfficientNet

EfficientNet was planned in view of the ConvNets neural architectural search and furthermore it has been created considering a versatile size baseline. This benchmark network is created by utilizing a neural architecture search with multiple objective that streamlines accuracy and FLOPS. The optimization goal for this method was $ACC(M) \times [FLOPS(m)/T]^w$, where $w = -0.07$ is a hyper-parameter that controls the trade-off between accuracy and FLOPS. $ACC(m)$ and $FLOPS(m)$ denote the accuracy and FLOPS of model (m), the target FLOPS is T, and $ACC(m)$ is the optimization goal. The versatile estimated EfficientNet model can be increased really, astounding cutting edge

precision with a significant degree less boundaries and FLOPS, on ImageNet and other transfer learning datasets [39].

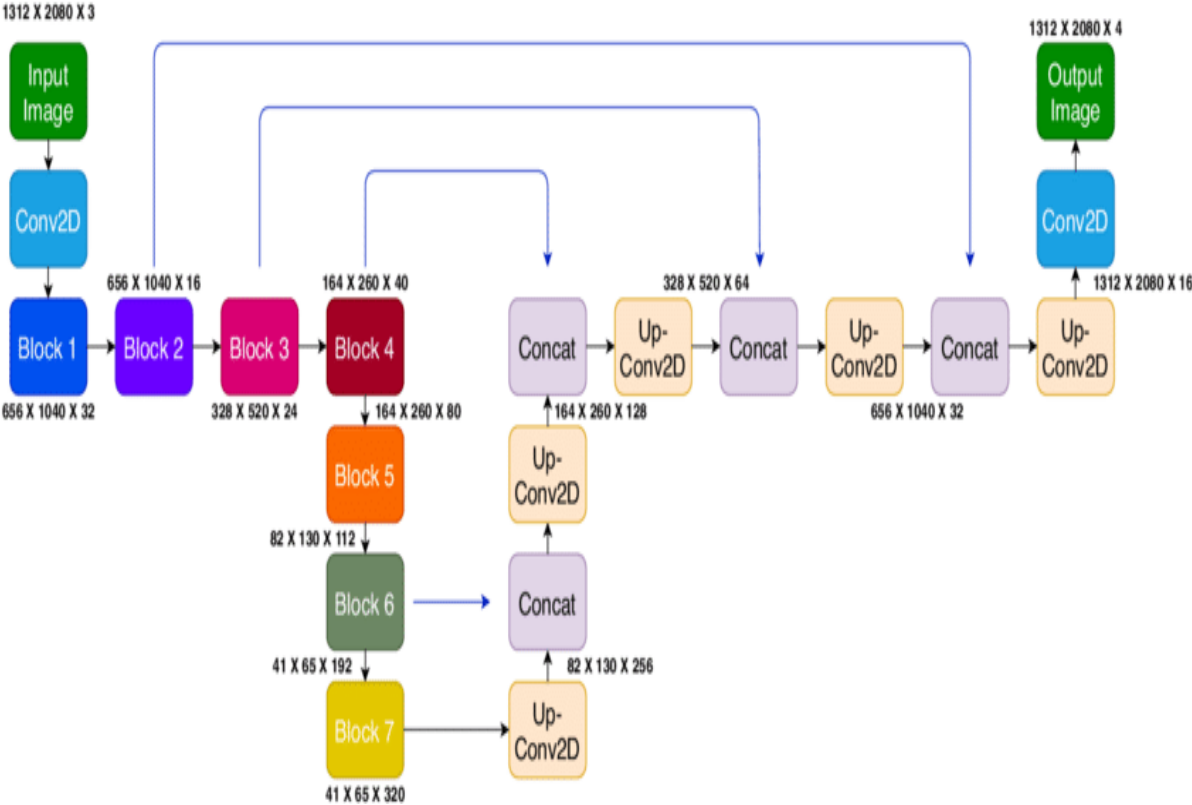


Figure 3.14 - Architecture of EfficientNet [52]

3.2.12 ConvNeXtXLarge

ConvNeXtXLarge developed considering the standard ConvNet modules, contend well with the transformers in the terms of precision, adaptability and strength across every one of the significant benchmarks. ConvNext maintains the efficiency of standard ConvNets and is extremely simple to implement due to its fully convolutional nature for both training and testing [40].

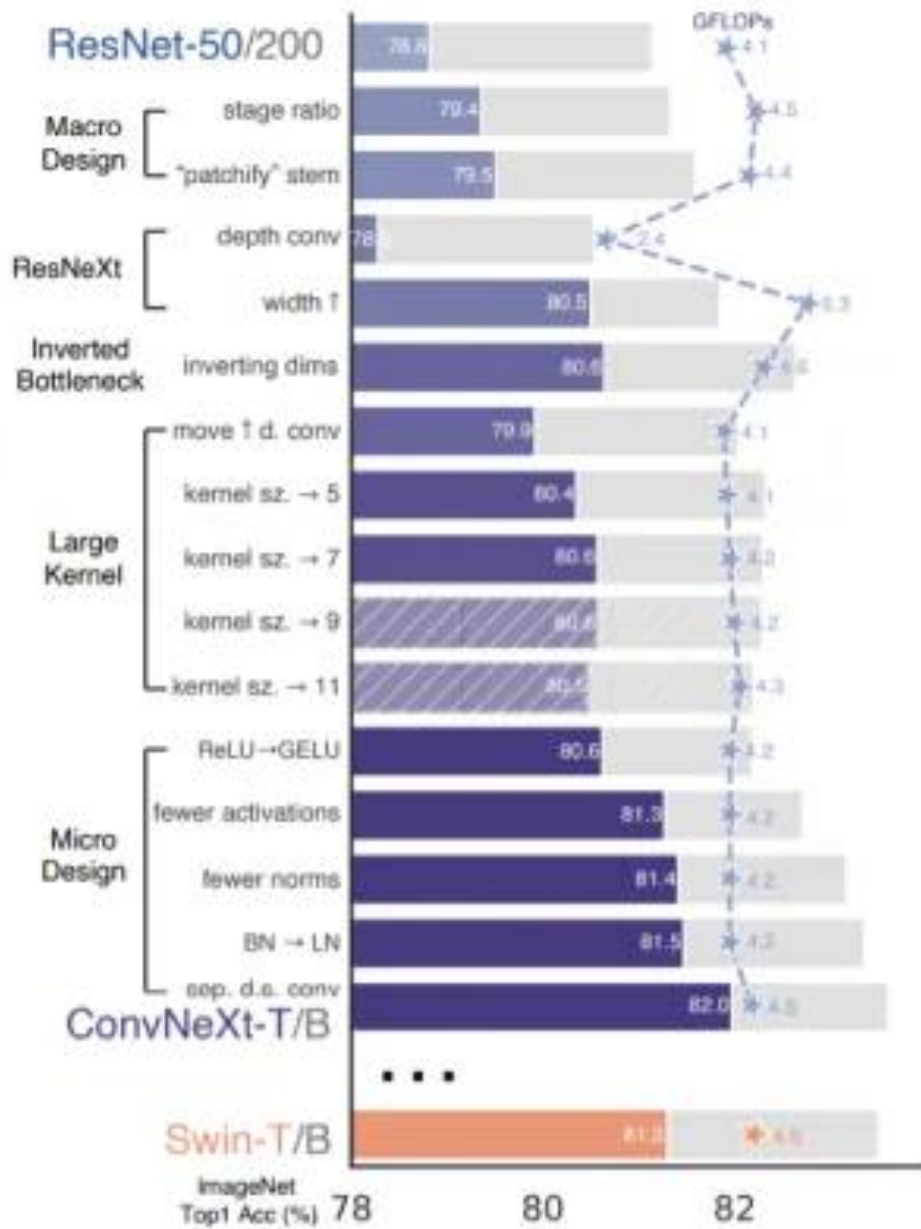


Figure 3.15 - Design of ConvNeXt [53]

3.3 CNN-based Transfer Learning Techniques

Machine learning has a technique called transfer learning where a pre-trained model is used for the training of similar kinds of tasks in a new model. It reduces computational power and costs of the model as the model does not have to start from scratch to train it.

Traditional ML VS Transfer Learning

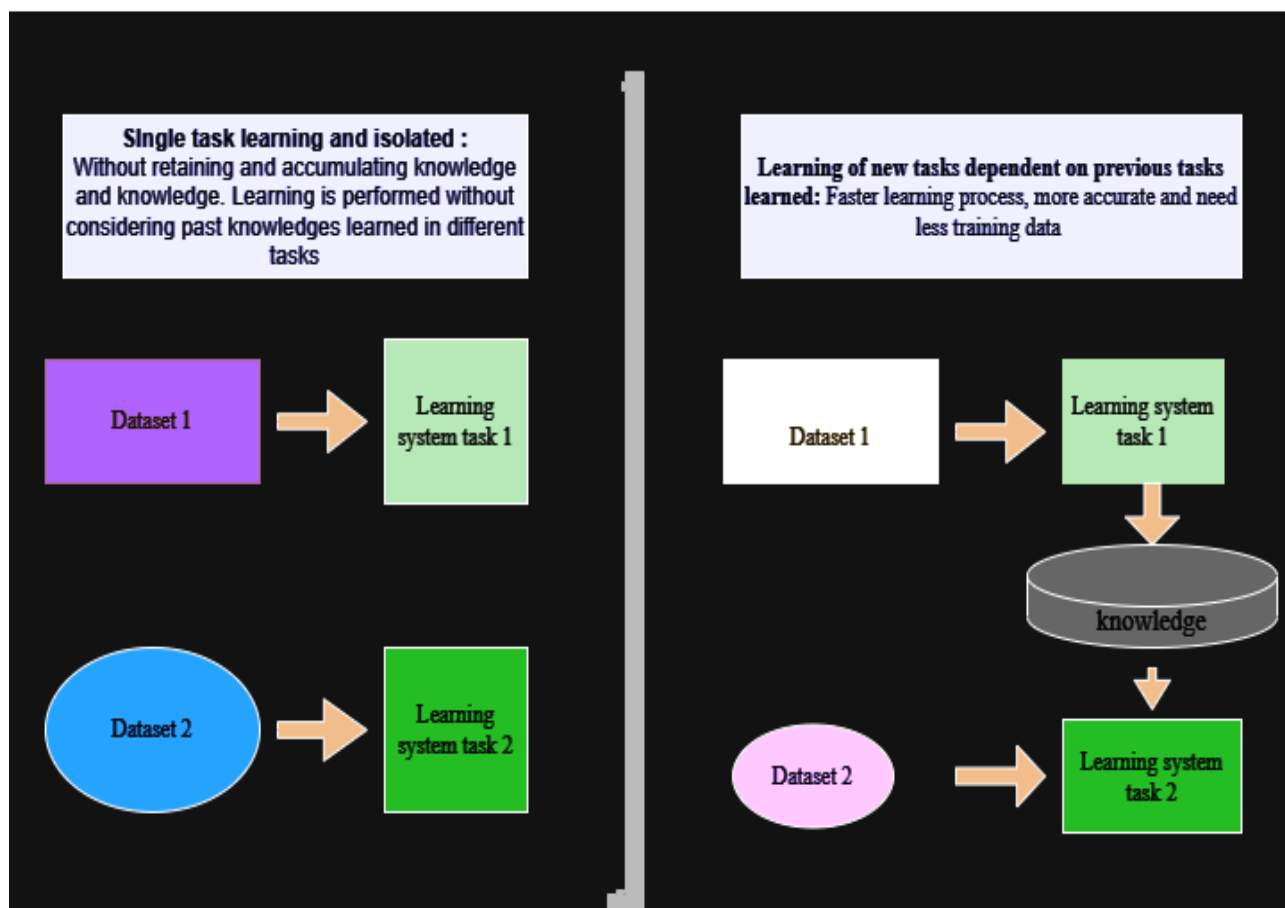


Figure 3.16: Working style of traditional machine learning and transfer learning

There are multiple CNN architectures which have practical applications for different kinds of image classifying tasks. CNN architecture uses a variety of image classification tasks as per the requirement. An architecture can be used in multiple ways to classify an image depending on the applications. The parameters of the architecture can be used with its characteristics for other new datasets. The weights are acted as the goal of CNN model which are changed according to the ideal results and afterward administered model preparation is finished on the test information [54]. Each CNN layer removes the significant picture data portraying another info picture portrayal for the next layer of its architecture.

Table - 3.1 Comparison of Different CNN Models

Model	Depth(Layers)	No. of Parameters (millions)
ResNet50	177	25.6
InceptionResNetV2	572	55.9
Xception	81	22.9
NasNet	1041	89.0
VGG19	26	143.7
MobileNet	88	4.3
DenseNet	121	8.1
ConvNeXtXLarge	296	350.1
EfficientNetV2L	1029	119.1

In this project work multiple transfer learning techniques used in the CNN based neural network architecture to collect the features for comparing mean square error. The modern transfer learning techniques use the information gained from the tasks done previously for learning purposes and categorizing the data instead of traditional transfer learning techniques.

We have used multiple transfer learning techniques on the same datasets to get the best mean square errors for crowd counting anomaly detection methods. This method performs prediction tasks for different testing samples using the transfer learning techniques like Xception, NasNet, VGG19, InceptionResNetV2, MobileNet, DenseNet, ConvNeXtXLarge, ResNet50, EfficientNetV2L.

Eliminating the background, extracting the features, and improving the occlusion factors to improve the accuracy in the crowd counting is the focus of this paper. AI approaches give the adaptability to foresee the future occasions utilizing the previous gaining from the label set of images. For the measurement of the difference in error in between the predicted and actual output machine learning model can be used for the prediction of the output.

CHAPTER 4

EXPERIMENT AND RESULTS

4.1 Dataset for the Experiment

In this project work used the publicly available dataset for training models to develop an advanced crowd counting method. The webcam installed in a mall used for the purpose of the experiment consists of RGB images of 2000 frames. The images were preprocessed for the training purposes and the dataset used to train the system partitioned into training and testing on 70-30 percent criteria. From the total images of 2000, 1600 used for training and 400 used for validation purposes. The parameters of concerned experiment include batch size, batch learning rate, epoch sizes. The batch sizes of the images used is 64. We have used 50 epochs for training each model.

Table - 4.1 Hyperparameters of CNN models

Parameters	Values
Batch Size	64
Batch earning rate	0.0010
Epoch size	50

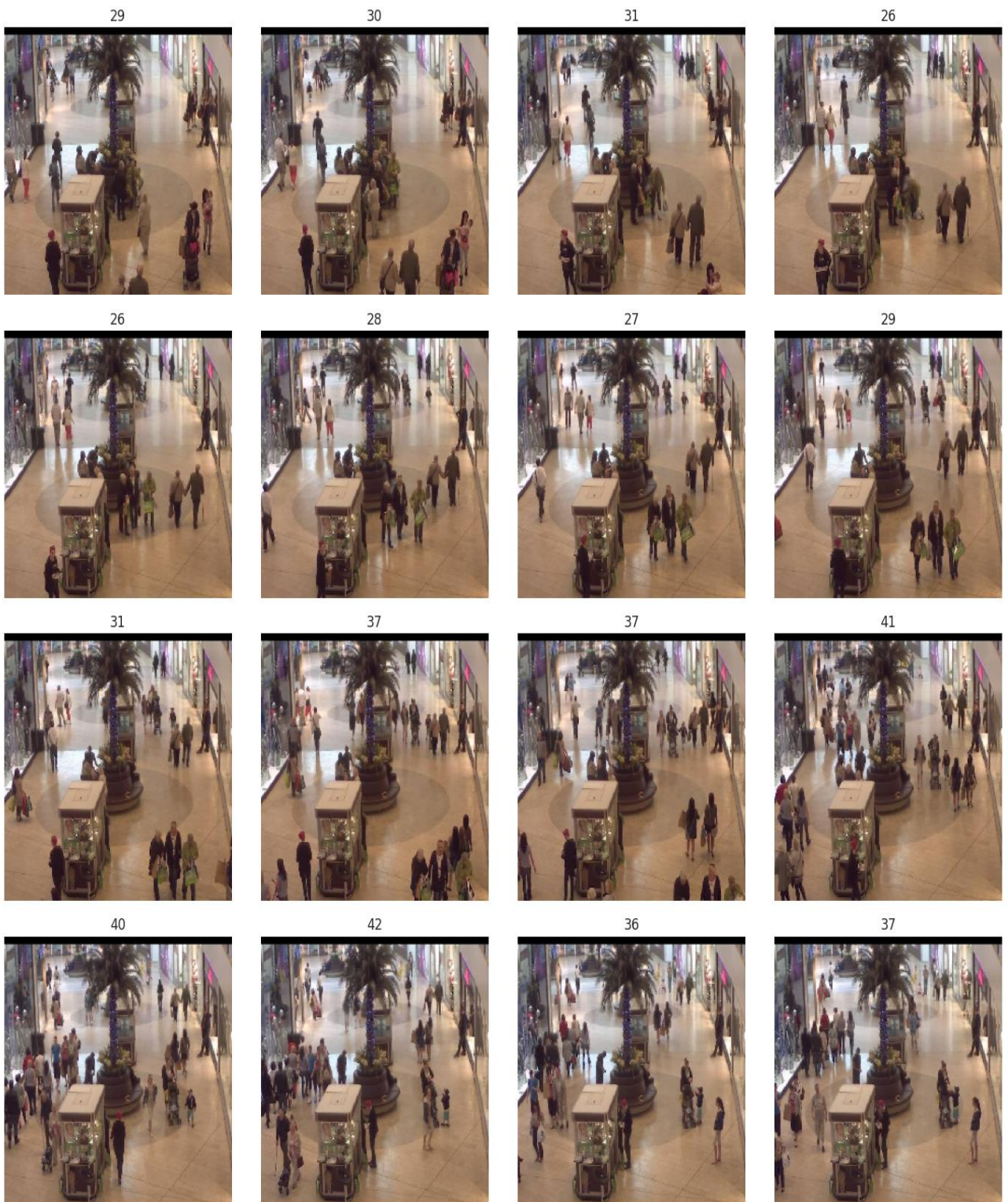


Figure – 4.1: Sample images of training dataset

Table - 4.2 Sample of the Dataset Labels

id	count
1	35
2	41
3	41
4	44
5	41
6	41
7	35
8	36
9	27
10	24
11	16
12	22
13	23
14	25
15	15
16	16
17	15
18	25
19	31
20	25
21	24
22	26
23	23
24	23
25	22
26	23
27	21
28	23
29	26
30	27

4.2 Methods Used for Performance Analysis

To improve the crowd counting accuracy authors used multiple transfer learning techniques to compare with the mean square error and Pearson r. Transfer learning used CNN models are trained on huge image datasets. In transfer learning it is possible to transfer information gained from previous tasks to improve performance on coming tasks. It benefits on saving time, improving performance, and requiring less data compared to training models from scratch. For the project work the purpose of using the transfer learning technique was to minimize the errors and increase the counting accuracy. The InceptionResNetV2 model gets the largest mean square error figure compared to other techniques whereas VGG19 performed best among the models but has the largest pearson r coefficient. The InceptionResNet and VGG19 both models trained on ImageNet dataset. ImageNet dataset has more than 14 million hand-annotated images which are categorized in more than 20,000 categories.

The expression use for the evaluation of the mean square error is:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{C}_i - C_i)^2$$

No of testing images = n

Actual number of human count in i^{th} = C_i

Count of the people estimated by the transfer learning model = \hat{C}_i

There is a need for two variables for measuring the linear relationship with the Pearson's correlation coefficient. The values of this parameter vary from -1 to +1. The positive worth demonstrates the propensity of one variable to increment or decline with deference of another variable though the negative worth sets that the rising worth of one variable outcomes in the abatement of another variable.

The equation use for the calculation of the Pearson's Correlation coefficient is:

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}}$$

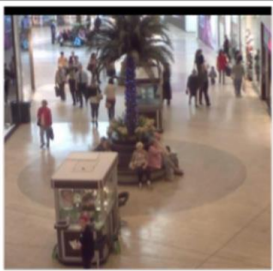
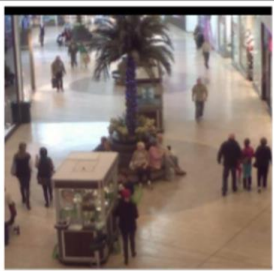
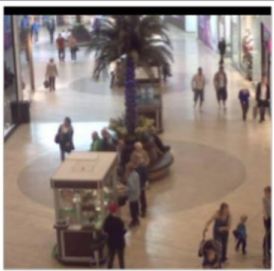
Pearson's Correlation Coefficient = r , Number of paired samples = n ,
 Actual count of the people = X , Estimated count of the people = Y .

4.3 Analysis of Results

Using the transfer learning techniques in this project work with the CNN based models it gives the predicted count based on original count from the training images almost better than expected.

From the original images with respect of original count using transfer learning techniques predicted count of the people given an idea of the working ability of these CNN models. With respect of the original count of 37, 28, 30 transfer learning technique ResNet50 has given predicted count 33, 29 and 30, InceptionResNetV2 has given 32, 32, 35, Xception has given 35,32,33, NasNet 31, 33, and 34, VGG19 36, 28,31, MobileNet 34, 31,34, DenseNet has given 35, 31,33, ConvNeXtXLarge 34,32,32 and EfficientNet has given 33, 32 and 34. It is observable that VGG19, MobileNet and DenseNet has performed better compared to other transfer learning techniques. VGG19 is the best performer among all other CNN models.

Table – 4.3 Predicted count with respect of Actual Count

Original Image				
Original Count		37	28	30
Predicted Count	ResNet50	33	29	34
	InceptionResNetV2	32	32	35
	Xception	35	32	33
	NasNet	31	33	34
	VGG19	36	28	31
	MobileNet	34	31	34
	DenseNet	35	31	33
	ConvNeXtXLarge	34	32	32
	EfficientNetV2L	33	32	34

In this project work comparison of nine transfer learning techniques using the parameters like MSE and Pearson r used to check the best efficient transfer learning technique for crowd counting. The Quantitative analysis among these transfer learning techniques will give an idea how VGG19 can be used for developing a CNN based crowd counting model. VGG19 performed best among all other CNN models for predicting the counting of the crowds.

Based on the comparing among the performance of the transfer learning technique based on CNN model using the parameters like mean square error and Pearson r also indicating the best working ability of the VGG19. Whereas models like ResNet50 have MSE and Pearson r 15.2 and 0.8, InceptionResNetV2 has 32.9 and 0.7, Xception has 11.6 and 0.9, NasNet has 31.8 and 0.6, MobileNet has 13.1 and 0.8, DenseNet has 12.6 and 0.9, ConvNeXtXLarge has 13.1 and 0.8 and EfficientNetV2L has 16 and 0.8, but VGG19 has 6.3 and 0.9. VGG19’s low mean error keeps VGG19 ahead than others to use for developing a model for enhancing crowd counting accuracy.

Table - 4.4 Performance of CNN Based Transfer Learning Techniques

Model	MSE	Pearson r
ResNet50	15.2	0.8
InceptionResNetV2	32.9	0.7
Xception	11.6	0.9
NasNet	31.8	0.6
VGG19	6.3	0.9
MobileNet	13.1	0.8
DenseNet	12.6	0.9
ConvNeXtXLarge	13.1	0.8
EfficientNetV2L	16	0.8

CHAPTER - 5

CONCLUSION

Multiple factors such as population growth, global urbanization, effective information dissemination, improved transportation, etc., are some of the events the frequency of crowded situations is rising worldwide. Because of the intricacy of these reasonable peculiarities, utilizing just ordinary information alone may not be ideal for abnormality recognition. The project work has discussed various technological advancements for managing large crowds. However, most of the technologies, particularly transfer learning methods and datasets, have only been tested in limited environments, and it has yet to be proven that they are effective in an integrated crowd management framework for large crowds. Counting crowds accurately is a live problem for many events all over the world. Transfer learning-based model CNN architectures can be utilized for extracting features and developing a better system for detecting anomalies in moving images. Many powerful applications of the CNN models have shown outstanding performance in different fields. Our results show how we can implement the best available transfer techniques to enhance the crowd counting. VGG19 can be implemented to develop a powerful artificial intelligent model for surveilling upon the crowd's movement as well as objects. Only with the 6.3 mean square error and 0.9 Pearson r, it is possible to improve more and utilize this as a working model. The future aim is to focus on reducing errors and developing a smooth crowd counting system.

REFERENCES

- [1] D. Sharma, A. P. Bhondekar, A. K. Shukla, and C. Ghanshyam, "A review on technological advancements in crowd management," *J Ambient Intell Human Comput*, vol. 9, no. 3, pp. 485–495, Jun. 2018, doi: [10.1007/s12652-016-0432-x](https://doi.org/10.1007/s12652-016-0432-x).
- [2] "From Morbi tragedy to Kumbh Mela stampede, a look at major disasters in India," India Today. Accessed: Mar. 06, 2024. [Online]. Available: <https://www.indiatoday.in/india/story/morbi-bridge-collapse-gujarat-list-of-major-disasters-at-public-places-in-india-2291541-2022-10-31>
- [3] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 21–37. doi: [10.1007/978-3-319-46448-0_2](https://doi.org/10.1007/978-3-319-46448-0_2).
- [4] "What Is Image Segmentation? | IBM." Accessed: Jun. 03, 2024. [Online]. Available: <https://www.ibm.com/topics/image-segmentation>
- [5] M. Soh, "Learning CNN-LSTM Architectures for Image Caption Generation". 2016
<http://cs224d.stanford.edu/reports/msoh.pdf>
- [6] T. Pang, P. Li, and L. Zhao, "A survey on automatic generation of medical imaging reports based on deep learning," *BioMed Eng OnLine*, vol. 22, no. 1, p. 48, May 2023, doi: [10.1186/s12938-023-01113-y](https://doi.org/10.1186/s12938-023-01113-y).
- [7] K. Doshi, "Audio Deep Learning Made Simple: Sound Classification, step-by-step," Medium. Accessed: Jun. 05, 2024. [Online]. Available: <https://towardsdatascience.com/audio-deep-learning-made-simple-sound-classification-step-by-step-cebc936bbe5>
- [8] "Machine Learning for Synthetic Data Generation: A Review." Accessed: Jun. 05, 2024. [Online]. Available: <https://arxiv.org/html/2302.04062v6/#bib.bib4>
- [9] Owaidah, A., et al. "Review of modelling and simulating crowds at mass gathering events: Hajj as a case study," in *Journal of Artificial Societies and Social Simulation*, vol. 22, no. 2, 2019. Accessed: Mar. 06, 2024. [Online]. Available: <https://www.jasss.org/22/2/9.html>

- [10] K. Doshi and Y. Yilmaz, “Continual Learning for Anomaly Detection in Surveillance Videos,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 1025–1034. doi: [10.1109/CVPRW50498.2020.00135](https://doi.org/10.1109/CVPRW50498.2020.00135).
- [11] A. B. Chan and N. Vasconcelos, “Bayesian Poisson regression for crowd counting,” in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 545–551. doi: [10.1109/ICCV.2009.5459191](https://doi.org/10.1109/ICCV.2009.5459191).
- [12] U. Singh, J.-F. Determe, F. Horlin, and P. De Doncker, “Crowd Monitoring: State-of-the-Art and Future Directions,” *IETE Technical Review*, vol. 38, no. 6, pp. 578–594, Nov. 2021, doi: [10.1080/02564602.2020.1803152](https://doi.org/10.1080/02564602.2020.1803152).
- [13] X. Zhang, S. Yang, Y. Y. Tang, and W. Zhang, “A thermodynamics-inspired feature for anomaly detection on crowd motions in surveillance videos,” *Multimed Tools Appl*, vol. 75, no. 14, pp. 8799–8826, Jul. 2016, doi: [10.1007/s11042-015-3101-8](https://doi.org/10.1007/s11042-015-3101-8).
- [14] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, “Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network,” *Sensors (Basel)*, vol. 19, no. 11, p. 2472, May 2019, doi: [10.3390/s19112472](https://doi.org/10.3390/s19112472).
- [15] M. Bortnikov, A. Khan, A. M. Khattak, and M. Ahmad, “Accident Recognition via 3D CNNs for Automated Traffic Monitoring in Smart Cities,” in *Advances in Computer Vision*, K. Arai and S. Kapoor, Eds., in *Advances in Intelligent Systems and Computing*. Cham: Springer International Publishing, 2020, pp. 256–264. doi: [10.1007/978-3-030-17798-0_22](https://doi.org/10.1007/978-3-030-17798-0_22).
- [16] Z. hui, X. yaohua, M. lu, and F. Jiansheng, “Vision-based real-time traffic accident detection,” in *Proceeding of the 11th World Congress on Intelligent Control and Automation*, Jun. 2014, pp. 1035–1038. doi: [10.1109/WCICA.2014.7052859](https://doi.org/10.1109/WCICA.2014.7052859).
- [17] Y.-C. Li, R.-S. Jia, Y.-X. Hu, D.-N. Han, and H.-M. Sun, “Crowd density estimation based on multi scale features fusion network with reverse attention mechanism,” *Appl Intell*, vol. 52, no. 11, pp. 13097–13113, Sep. 2022, doi: [10.1007/s10489-022-03187-y](https://doi.org/10.1007/s10489-022-03187-y).
- [18] W. Sultani, C. Chen, and M. Shah, “Real-world Anomaly Detection in Surveillance Videos.” arXiv, Feb. 14, 2019. doi: [10.48550/arXiv.1801.04264](https://doi.org/10.48550/arXiv.1801.04264).
- [19] T. Ohgushi, K. Horiguchi, and M. Yamanaka, “Road Obstacle Detection Method Based on an Autoencoder with Semantic Segmentation,” H. Ishikawa, C.-L. Liu, T. Pajdla, and J. Shi, Eds., in

Lecture Notes in Computer Science, vol. 12627. Cham: Springer International Publishing, 2021, pp. 223–238. doi: [10.1007/978-3-030-69544-6_14](https://doi.org/10.1007/978-3-030-69544-6_14).

[20] B. Varona, A. Monteserin, and A. Teyseyre, “A deep learning approach to automatic road surface monitoring and pothole detection,” *Pers Ubiquit Comput*, vol. 24, no. 4, pp. 519–534, Aug. 2020, doi: [10.1007/s00779-019-01234-z](https://doi.org/10.1007/s00779-019-01234-z).

[21] A. Al-Dhamari, R. Sudirman, and N. H. Mahmood, “Transfer Deep Learning Along With Binary Support Vector Machine for Abnormal Behavior Detection,” *IEEE Access*, vol. 8, pp. 61085–61095, 2020, doi: [10.1109/ACCESS.2020.2982906](https://doi.org/10.1109/ACCESS.2020.2982906).

[22] M. Sabokrou, M. Fayyaz, M. Fathy, Zahra. Moayed, and R. Klette, “Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes,” *Computer Vision and Image Understanding*, vol. 172, pp. 88–97, Jul. 2018, doi: [10.1016/j.cviu.2018.02.006](https://doi.org/10.1016/j.cviu.2018.02.006).

[23] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, “CNN-based Density Estimation and Crowd Counting: A Survey.” arXiv, Mar. 28, 2020. doi: [10.48550/arXiv.2003.12783](https://doi.org/10.48550/arXiv.2003.12783).

[24] K. Xu *et al.*, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.” arXiv, Apr. 19, 2016. doi: [10.48550/arXiv.1502.03044](https://doi.org/10.48550/arXiv.1502.03044).

[25] L. Gu, C. Pang, Y. Zheng, C. Lyu, and L. Lyu, “Context-aware pyramid attention network for crowd counting,” *Appl Intell*, vol. 52, no. 6, pp. 6164–6180, Apr. 2022, doi: [10.1007/s10489-021-02639-1](https://doi.org/10.1007/s10489-021-02639-1).

[26] C. C. Loy, S. Gong, and T. Xiang “From Semi-Supervised to Transfer Counting of Crowds”, In Proceedings of IEEE International Conference on Computer Vision, (ICCV) pp. 2256-2263, 2013

K. Chen, S. Gong, T. Xiang, and C. C. Loy “Cumulative Attribute Space for Age and Crowd Density Estimation”, in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2467-2474, 2013 (CVPR, Oral)

C. C. Loy, K. Chen, S. Gong, T. Xiang “Crowd Counting and Profiling: Methodology and Evaluation” in S. Ali, K. Nishino, D. Manocha, and M. Shah (Eds.), Modeling, Simulation and Visual Analysis of Crowds, Springer, vol. 11, pp. 347-382, 2013

K. Chen, C. C. Loy, S. Gong, and T. Xiang “Feature Mining for Localized Crowd Counting” British Machine Vision Conference, 2012 (BMVC)

Official link: http://personal.ie.cuhk.edu.hk/~ccloy/downloads_mall_dataset.html

- [27] C. Bhardwaj, S. Jain, and M. Sood, “Transfer learning based robust automatic detection system for diabetic retinopathy grading,” *Neural Comput & Applic*, vol. 33, no. 20, pp. 13999–14019, Oct. 2021, doi: [10.1007/s00521-021-06042-2](https://doi.org/10.1007/s00521-021-06042-2).
- [28] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines | Proceedings of the 27th International Conference on International Conference on Machine Learning,” Guide Proceedings. Accessed: Jun. 05, 2024. [Online]. Available: <https://dl.acm.org/doi/10.5555/3104322.3104425>
- [29] Moscoso Alcantara EA, Bong MD, Saito T. “Structural Response Prediction for Damage Identification Using Wavelet Spectra in Convolutional Neural Network.” *Sensors (Basel)*. 2021 Oct 13;21(20):6795. doi: 10.3390/s21206795. PMID: 34696008; PMCID: PMC8539720.
- [30] Y. LeCun *et al.*, “Handwritten Digit Recognition with a Back-Propagation Network,” in *Advances in Neural Information Processing Systems*, Morgan-Kaufmann, 1989. Accessed: Mar. 07, 2024. [Online]. Available: <https://proceedings.neurips.cc/paper/1989/hash/53c3bce66e43be4f209556518c2fcb54-Abstract.html>
- [31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2012. Accessed: Mar. 07, 2024. [Online]. Available: <https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>
- [32] C. Szegedy *et al.*, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 1–9. doi: [10.1109/CVPR.2015.7298594](https://doi.org/10.1109/CVPR.2015.7298594).
- [33] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition.” arXiv, Apr. 10, 2015. doi: [10.48550/arXiv.1409.1556](https://doi.org/10.48550/arXiv.1409.1556).
- [34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).

- [35] F. Chollet, "Xception: Deep Learning with Depth Wise Separable Convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 1800–1807. doi: [10.1109/CVPR.2017.195](https://doi.org/10.1109/CVPR.2017.195).
- [36] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 8697–8710. doi: [10.1109/CVPR.2018.00907](https://doi.org/10.1109/CVPR.2018.00907).
- [37] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." arXiv, Apr. 16, 2017. Accessed: Feb. 27, 2024. [Online]. Available: <http://arxiv.org/abs/1704.04861>
- [38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. "Densely connected convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700-4708. 2017. Accessed: Feb. 27, 2024. [Online].
- [39] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." In *International conference on machine learning*, pp. 6105-6114. PMLR, 2019. Accessed: Feb. 29, 2024. [Online].
- [40] Z. Liu, H. Mao, C.-Y Wu, C. Feichtenhofer, T. Darrell, and S. Xie. "A convnet for the 2020s." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976-11986. 2022. Accessed: Mar. 07, 2024. [Online].
- [41] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1. 2017. Accessed: Mar. 07, 2024. [Online].
- [42] T. Ghazal, S. Munir, S. Abbas, A. Athar, H. Alrababah, and M. Khan, "Early Detection of Autism in Children Using Transfer Learning," *IASC*, vol. 36, no. 1, pp. 11–22, 2022, doi: [10.32604/iasc.2023.030125](https://doi.org/10.32604/iasc.2023.030125).
- [43] A. BRITAL, "GoogLeNet CNN Architecture Explained (Inception V1).," Medium. Accessed: May 27, 2024. [Online]. Available: <https://medium.com/@AnasBrital98/googlenet-cnn-architecture-explained-inception-v1-225ae02513fd>

- [44] "Understanding ResNet-50 in Depth: Architecture, Skip Connections, and Advantages Over Other Networks - Wisdom ML." Accessed: May 27, 2024. [Online]. Available: <https://wisdomml.in/understanding-resnet-50-in-depth-architecture-skip-connections-and-advantages-over-other-networks/>
- [45] Nimai Chand Das Adhikari, "Infection Severity Detection of CoVID19 from X-Rays and CT Scans Using Artificial Intelligence," in *IJC*, Vol. 38. No. 1, pp. 73-92 , 2020.
- [46] A. BRITAL, "Inception V3 CNN Architecture Explained.," Medium. Accessed: May 27, 2024. [Online]. Available: <https://medium.com/@AnasBrital98/inception-v3-cnn-architecture-explained-691cfb7bba08>
- [47] Li, L., et al. "Real-time one-shot learning gesture recognition based on lightweight 3D Inception-ResNet with separable convolutions," in *Pattern Analysis and Applications*, vol. 24, pp. 1-20, 2021.
- [48] Abdalkarim Mohtasib, undefined., et al. "Neural Task Success Classifiers for Robotic Manipulation from Few Real Demonstrations," in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, 2021.
- [49] Munadi, K., et al. "A Deep Learning Method for Early Detection of Diabetic Foot Using Decision Fusion and Thermal Images," in *Applied Sciences*, vol. 12, pp. 7524, 2022.
- [50] Srinivasan, K., et al. "Performance Comparison of Deep CNN Models for Detecting Driver's Distraction," in *Cmc -Tech Science Press-*, vol. 68, pp. 4109-4124, 2021.
- [51] A. Aytekin, V. Mençik, C. Budak, "AM-DSB MODULATION DETECTION AMONG SIGNALS MODULATED WITH 26 DIFFERENT MODULATION TECHNIQUES WITH MobileNet ARCHITECTURE," in *International Informatics Congress*, 2022.
- [52] T. Ahmed, N. Sabab. "Classification and Understanding of Cloud Structures via Satellite Images with EfficientUNet," in *SN Computer Science*, vol. 3, 2022.

[53] Singh, “ConvNext: The Return Of Convolution Networks,” Augmented Startups. Accessed: May 27, 2024. [Online]. Available: <https://medium.com/augmented-startups/convnext-the-return-of-convolution-networks-e70cbe8dabcc>

[54] S. Mohammadian, A. Karsaz, and Y. M. Roshan, “Comparative Study of Fine-Tuning of Pre-Trained Convolutional Neural Networks for Diabetic Retinopathy Screening,” in *2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME)*, pp. 1–6. Nov. 2017, doi: [10.1109/ICBME.2017.8430269](https://doi.org/10.1109/ICBME.2017.8430269).

PUBLICATIONS

A. Roy, N. Jain, V. Baghel, “Estimation of Abnormal Crowd Density Using Transfer Learning Techniques” - **accepted** in the 15th International IEEE Conference on Computing, Communication And Networking Technologies (ICCCNT) organized by IIT Mandi, Himachal Pradesh.

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT
PLAGIARISM VERIFICATION REPORT

Date: 6/6/2024

Type of Document (Tick): PhD Thesis M.Tech/M.Sc. Dissertation B.Tech./B.Sc./BBA/Other

Name: Amit Roy Department: ECE Enrolment No 225042002

Contact No. 8981301503 E-mail. 225042002@JuitSolon.in

Name of the Supervisor: Dr. Nishant Jain & Dr. Vikas Baghel

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): ANOMALY DETECTION IN SURVEILLANCE VIDEOS

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

- Total No. of Pages = 62
- Total No. of Preliminary pages = 18
- Total No. of pages accommodate bibliography/references = 7

Amit
(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at 10 (%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

Nishant
6/6/24
(Signature of Guide/Supervisor)

Jain
07/06/2024
Signature of HOD

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Abstract & Chapters Details	
<u>06th/06/2024</u>	<ul style="list-style-type: none"> • All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String 	<u>08%</u>	Word Counts	<u>7406</u>
Report Generated on			Character Counts	<u>40,860</u>
<u>07th/06/2024</u>			Page counts	<u>42</u>
		Submission ID	File Size	<u>6.76 M</u>
		<u>2396802608</u>		

[Signature]
Checked by 07/06/24
Name & Signature

[Signature]
07-06-24
Librarian

Please send your complete Thesis/Report in (PDF) & DOC (Word File) through your Supervisor/Guide at plagcheck.juit@gmail.com

LIBRARIAN
 RESOURCE CENTRE
 Jaypee University of Information Technology
 Waknaghat, Distt. Solan (Himachal Pradesh)
 Pin Code - 173234